Mitochondrial Haplogrouping and Short Tandem Repeat Analyses in Anthropological Research using Next-Generation Sequencing Technologies

By
Melody D. Ratliff Wood
M.A., University of Montana, Missoula, MT, 2014
B.A., University of Tennessee, Knoxville, TN, 2012

Submitted to the graduate degree program in Biological Anthropology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chair: Michael H. Crawford

_____

Jennifer Raff

_____

Dennis O'Rourke

_____

Maria Orive

_____

Charla Marshall

Date Defended: May 10, 2018

The dissertation committee for Melody D. Ratliff Wood certifies that this is the approved version of the following dissertation:

Mitochondrial Haplogrouping and Short Tandem Repeat Analyses in Anthropological Research using Next-Generation Sequencing Technologies

Chair: Michael H. Crawford

Date Approved: June 11, 2018

Abstract

The field of anthropological genetics aims to reveal, characterize, and understand the biological diversity of modern and ancient human populations. This goal is achieved by analyzing different regions of the autosomes, sex chromosomes, and mitochondrial genome. The last decade has introduced a new wave of technologies known as next-generation sequencing (NGS) technologies with high throughput and increased data output. NGS has been employed in the medical and forensic fields but is slow to take hold in anthropological genetics. This work demonstrates the utility of NGS to answer anthropological questions and genetically characterize populations. The accuracy of mitochondrial haplogrouping using smaller ranges of the mitogenome was assessed. When using less than the full mitogenome, haplogrouping was accurate for 95% of samples. Using only the control region, 50% of samples were precisely haplogrouped and 82% of Native American haplogroups were distinguishable from Asian haplogroups. Examining autosomal and Y-chromosome STRs, nine loci exhibited increased sequence-based allelic diversity. Five loci (D2S441, D7S820, vWA, DYS392, DYS635) demonstrated statistical differences in the frequency distributions of length-based and sequence-based alleles for Native American and Asian samples; two of these loci (vWA and DYS635) demonstrated higher significance levels when using sequence-based alleles. One locus (D2S1338) demonstrated statistical differences in the sequence-based alleles alone. This indicates the D2S1338, vWA, and DYS635 loci are populationally informative using sequence-based alleles obtained by NGS. These are some of the fundamental areas in which anthropological genetics can advance using next-generation sequencing technologies.

Acknowledgements

I would like to thank my committee, Jennifer Raff, Dennis O'Rourke, Charla Marshall, and Maria Orive, and my adviser, Michael Crawford, for providing me with their expertise throughout this research project and over these past four years. Thank you for helping me complete my degree goals and agreeing to take me on as a graduate student. I would like to thank Michelle Peck, Joseph Ring, Erin Gorden, Jennifer Higginbotham, and particularly Charla Marshall and Kim Sturk-Andreaggi at AFMES-AFDIL for laboratory assistance, writing feedback, and for continual help along the way. AFDIL provided me with such a great knowledge base of NGS and a wonderful research opportunity for which I am forever grateful. I would like to thank Justin Tackney at KU for help with phylogenetic construction and would like to thank all collaborators that contributed samples to make this research possible. I would also like to thank the graduate students in the Laboratories of Biological Anthropology for assistance, academic discussion, and friendship over these past few years. Finally, I would not be able to complete this goal of mine without the support of my family, friends, and the unwavering support of my husband, Chance, who encouraged me to pursue this degree after I had already given up.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The field of anthropological genetics aims to reveal, characterize, and understand the biological diversity of modern and ancient human populations. This goal is achieved by analyzing different regions of the autosomes, sex chromosomes, and mitochondrial genome. Over the last few decades, advances in sequencing capabilities have continually improved. Specifically, the last decade has introduced a new wave of technologies known as next-generation sequencing (NGS) technologies. These technologies are also referred to as massively parallel sequencing because they are capable of sequencing millions of small DNA fragments in parallel, meaning, at the same time.

Next generation sequencing has expanded our ability to examine human genetic variability more than ever by providing substantially higher throughput and increased data output. However, these new technologies come with higher costs and increased laboratory procedures, delaying the adoption of such technologies in some fields of genetic research. Because NGS requires expensive laboratory equipment, consumables, and data management programs, its use was initially limited to well-funded laboratories. As these technologies have continued to advance and produce industry competition, prices have decreased drastically, making it more affordable for laboratories with less funding. Presently, the tradeoff between costs and utility have shifted; the value added from NGS-derived data is outweighing the costs in the eye of many laboratories.

NGS technologies were first routinely employed in medical research and clinical diagnostics for the screening of mutations in hundreds of loci to identify novel mutations and identify genetic disorders (Choi et al., 2009; Tucker et al., 2009; Kinde et al., 2011). Clinical NGS research has since expanded into several research interests including pharmacogenetics,

epigenetics, cancer genetics and the transmission of complex traits (Tucker et al., 2009; Neverov et al., 2010; Zeggini, 2011; Tzvetkov and von Ahsen, 2012; Park et al., 2013; Ezewudo and Zwick, 2013; Matullo et al., 2013; Suzuki and Greally, 2013; Guchelaar et al., 2014). Gradually, NGS technologies moved into the forensic realm to aid in human identification. Forensic identification traditional relied on capillary electrophoresis (CE) and length variant comparisons of autosomal and Y-chromosome short tandem repeats (STRs). However, with NGS, sequence data was obtained in addition to the length variants (Berglund et al., 2011), which provided a method to distinguish between mixed samples and increased likelihood ratios in identification (Ballantyne et al., 2010; Kayser & de Knijff, 2011; Bornman et al., 2012; Gelardi et al., 2014; Aly and Sabri, 2015; Caratti et al., 2015). In addition to STR profiling, NGS provides increased depth of coverage in the mitochondrial genome. Sanger-type sequencing can identify heteroplasmies when the minor allele is above 20% (Alvarez-Cubero et al., 2017) while NGS can identify heteroplamsies at a much more sensitive level, when the minor allele is below 5% (Holland et al., 2011; Just et al., 2015). This allows for the positive identification of heteroplasmies that may be missed using the Sanger-type sequencing (Coble et al., 2009; Sosa et al., 2012).

Researchers in anthropological genetics are slowly beginning to utilize NGS technologies as they become more affordable and the value of the added data is fully realized. Many of the first uses of NGS in anthropology were focused on notable, individual specimens, such as the Tyrolean Iceman, sequenced in 2008 (Ermini et al., 2008). At the same time, there was an increase in the number of mitogenome sequencing using NGS techniques seen with ancient samples to aid in the reconstruction of past human migrations and genetic structure (Cui et al., 2013; Veeramah and Hammer, 2014; Tackney et al., 2015; Llamas et al., 2016; Matisoo-Smith et

al., 2018). However, anthropological research involving modern populations has often overlooked such in-depth analyses until more recently. For example, only in 2016 were comparative, modern-day Alpine populations sequenced for full mitogenome data to compare with the ancient mitogenome of the famous Tyrolean Iceman (Coia et al., 2016). Today, several anthropological studies have begun to utilize NGS for the genetic characterization of modern populations (Lopopolo et al., 2016; Nagle et al., 2017; Neparáczki et al., 2017), though many studies continue to focus on smaller segments of mtDNA.

Sanger-type sequencing was first capable of identifying 15 to 200 nucleotides from the priming site using a single primer (Sanger et al., 1977) but has advanced over the years to enable reads up to 900 base pairs in length (Morozova and Marra, 2008; Heather and Chain, 2016). Thus, the Sanger sequencing method is limited technically with mtDNA as all 16,569 base pairs cannot be practically sequenced, particularly with minute samples (Parson et al., 2013). For this reason, only smaller, more informative regions are chosen for sequencing using Sanger methods (Yang et al., 2014). Focus is placed on the control region (CR), more specifically hypervariable (HV) segments 1 and 2. These are non-coding regions that are highly polymorphic, making them ideal for anthropological studies. NGS can significantly increase the regions of mtDNA to be sequenced, including the sequencing of the full mitogenome using only two long-range primer pairs.

This research will examine the performance of data output from next generation sequencing and its utilization in anthropological genetics for full mitogenome analyses. As many previous anthropological studies only utilized small segments of the mitogenome, inaccurate estimations of haplogroups may have occurred. The main question lies with how

often this happens. This research utilizes mitochondrial genomes to answer the following question:

1) How accurate are haplogroup assignments when only portions of the mitogenome are sequenced?

This is tested by examining various ranges of the mitogenome commonly used to make haplogroup assignments against haplogroup assignments made using the full mitogenome. Haplogrouping accuracy rates when using less than the full mitogenome is determined. The implications of these errors are discussed. Further, rates of precise haplogrouping when using less than the full mitogenome are determined.

Comparing results from complete mitogenome sequencing via NGS versus partial mitochondrial sequencing via Sanger sequencing methods, it has been noted that in some cases, alternative haplogroup assignments were produced. This is due to the exploration of sequence information outside of the control region. King and colleagues (2014) examined 283 samples using both whole mitogenome sequencing via NGS and partial mitochondrial sequencing Sanger methods. Approximately 3% of the samples changed haplogroup clade (e.g. G to L) when whole mitogenome data was reported from NGS versus HV 1 and 2 data reported from Sanger sequencing. Furthermore, 2% of these changed macrohaplogroups (i.e. L, M, N). Neparáczki et al. (2017) revealed several incorrect haplogroup assignments from HV data as compared to full mitogenome data using NGS during a recent reanalysis of ancient Hungarian samples. These results demonstrate how mitogenome sequencing versus only HV 1 and 2 can reveal crucial additional variants that can significantly improve the discriminatory power and assignment of mitochondrial haplogroups.

One example of such necessary discriminatory power is seen in mtDNA haplogroup B4b. There are two divergent branches from haplogroup B4b: B4b1 and B2. Haplogroup B2 is found exclusively among indigenous peoples of the Americas (Achilli et al., 2008; O'Rourke and Raff, 2010; Achili et al., 2013), while haplogroup B4b1 is only seen in Asia (Torroni et al., 1993a; Tanaka et al., 2004; Starikovskaya et al., 2005; Derenko et al., 2012). Many variants needed to distinguish haplogroup B2 from B4 are predominately located in the mtDNA coding region. Therefore, these variants are not detected if only the smaller HV segments of the mtDNA are examined, which can be problematic when trying to determine accurate haplogroups and ancestral affiliation. This research will examine if increased discriminatory power of ancestral identification is provided when using full mitogenome data. This will be demonstrated by the ability to discern Native American ancestry from Asian ancestry within haplogroup B. This research utilizes mitochondrial genomes to answer the following question:

2) Within Haplogroup B, how often are Native American subgroups distinguishable from Asian subgroups when only a portion of the mitogenome is sequenced?

I test this by examining several regions of mtDNA as compared to the full mitogenome. Frequencies of Native American haplogroup distinction from Asian haplogroups are calculated using increasing ranges of the mitogenome. Regions of the mitogenome necessary for distinguishing B2 from B4b are determined. Further, an alternative method for distinguishing B2 from B4b using only data from the control region is discussed.

In addition to mtDNA, anthropological genetics examines short tandem repeat (STR) loci of the sex chromosomes and autosomes. STRs are sequences of 2 to 6 base pairs that variably repeat, usually between 10 and 30 times, and are thought to result from slippage of strand pairing

during DNA replication (Dauber et al., 2012). In anthropological research, Y-chromosome STRs and autosomal STRs are commonly used to derive population information (Mitchell et al., 2006; Kim, et al., 2007; Nuñez et al., 2010; Young et al., 2011; Crawford and Beaty, 2013; He et al., 2017). STRs of the Y-chromosome (Y-STRs) are specifically utilized to determine paternally inherited haplogroups that can provide ancestral information of male lineages. Autosomal STRs (aSTRs) are used to provide population information that is not sex-specific, as autosomes are inherited from both the mother and the father.

Generally, STRs are obtained using more common typing methods that utilize capillary electrophoresis (CE) that provides information about the length of each repeat. The length of each repeat tends to vary widely between individuals, making them useful in forensic identification. However, they do display population structure that can be used in population genetic research. With newer NGS sequencing methods, not only can the number of repeats (i.e. allele length) be determined, but the actual nucleotide sequence is revealed.

Using next-generation sequencing of STRs has revealed there is significant variation within STR alleles that provide higher discriminatory power (Gelardi et al., 2014; Scheible et al., 2014). While two individuals share the same length of an allele (i.e. share the same number of repeats at a loci), they may differ in their nucleotide sequences. Recent research has suggested these intra-repeat variants, also known as isoalleles, may be non-randomly distributed across populations (Planz et al., 2012; Warshauer et al., 2015; Wang et al., 2017) and across all loci (Gettings et al., 2016; van der Gaag et al., 2016; Guo, 2017). Therefore, population characterization of STR isoalleles may be of interest not only for forensic identification likelihoods, but also for anthropological inferences. This research will examine the prevalence

of isoalleles from next-generation sequencing of autosomal and Y-chromosome STRs to answer the following question:

3) Are isoallelic frequencies distributed randomly or non-randomly across populations?

I test this by examining frequency distributions of sequence-based alleles compared to length-based alleles for Native American individuals and Asian individuals. If the frequency distributions of sequence-based alleles as compared to length-based alleles provides increased statistical differentiation between the two groups, then it would indicate isoalleles are distributed non-randomly across populations and can be useful for characterizing populations.

The addition of sequence data may only be informative if sequence-based alleles are distributed non-randomly across populations. If isoallelic frequencies vary significantly between populations or groups, then these can be useful in anthropological genetic research to better characterize populations. This research will examine the prevalence of isoalleles from next-generation sequencing of autosomal and Y-chromosome STRs to answer the following question:

4) Can isoalleles be populationally informative for use in anthropological genetics?

I test this by examining the frequency distributions of Native American individuals compared to Asian individuals for several STR loci. If tests of significance reveal statistically significant differences between groups using sequence-based alleles, this will indicate population structure exists among isoalleles and can be used in population genetic research as another method for characterizing populations.

The following work is divided into six additional chapters which aim to answer these proposed questions. Chapter 2 details next-generation sequencing technologies used in both

mitochondrial and STR analyses. Chapter 3 specifically details mtDNA analyses using next-generation sequencing technologies. Chapter 4 details the mitochondrial analyses of this dissertation including samples, methods, statistical analyses, results, and discussion. Chapter 5 specifically details autosomal and Y-chromosome STR analyses using next-generation sequencing technologies. Chapter 6 details the STR analyses of this dissertation including samples, methods, statistical analyses, results, and discussion. Chapter 7 summarizes the overall findings and describes the significance of this research.

Chapter 2: Next-Generation Sequencing Technologies

Early sequencing technologies, sometimes called first generation sequencing, include the early methods of sequencing DNA, prior to the high throughput methods of today. Sanger sequencing, or the Sanger dideoxynucleotide triphosphate (ddNTP) chain terminating method, was developed in 1977 and remained the most widely used method for many years and is still used today (Børsting and Morling, 2015; Liu et al., 2012). In Sanger sequencing, four ddNTPs, one for each base, are incorporated as DNA polymerase synthesizes a complementary strand. These ddNTPs differ from natural deoxynucleotides as they lack a 3' hydroxyl group necessary for base extension. Thus, the DNA chain is stopped once a terminator is incorporated into the DNA strand. Each terminator is labelled with a different fluorophore. Numerous copies of variable length fragments are generated in each reaction, all with a fluorophore attached to the end of each fragment.

Following this methodological step, capillaries are used to separate the DNA molecules by length. Electrophoresis through these capillaries separates each DNA fragment by size, differentiating between lengths of fragments that differ in length by only one nucleotide. The shorter fragments travel further through the capillary than longer fragments. As the DNA fragments are separated by length, each specific fluorophore attached to the end of each fragment can be determined: ddATP, ddGTP, ddCTP, or ddTTP (Shendure et al., 2008), revealing the DNA sequence of interest. The idea is that with so many variable length fragments, at least one fragment will incorporate a fluorophore at each base position, covering the entire region.

In the mid-2000s, a new suite of sequencing technologies was introduced. Next-generation sequencing (NGS) far surpassed traditional methods and revolutionized genomic research whereby large numbers of DNA sequences could be ascertained in a single reaction. In

the last decade, numerous NGS platforms have been established. Many of these use the similar chemistries to accomplish the goal of high throughout sequencing. Table 2.1 lists the most commonly used platforms and their specifications.

Table 2.1: Next-Generation Sequencing Platforms. Maximum output, read length, read per run, sequencing time, error rate, cost, and chemistry listed for each platform. Adapted from Kulski (2016). Information gathered from: (Li et al., 2005; Mardis, 2008; Metzker, 2010; Mardis, 2011; Lam et al., 2012; Liu et al., 2012; Gužvić, 2013; Kwong et al., 2015)

NGS platforms/company/ max output per run	**Read length per run (bp)**	**No. reads per run**	**Time (h or days)**	**Cost per 10^6 bases**	**Raw error rate (%)**	**Platform cost (USD approx.)**	**Chemistry**
First generation							
Sanger/Life Technologies/84 kb	800	1	2 h	2400	0.3	95,000	Dideoxy terminator
Second generation							
454 GS FLX+/Roche/0.7 Gb	700	$1x10^6$	24/48 h	10	1	500,000	Pyro-sequencing
GS Junior/Roche/ 70 Mb	500	$1x10^5$	18 h	9		100,000	Pyro-sequencing
HiSeq/Illumina/ 1500 Gb	2x150	$5x10^9$	27/240 h	0.1	0.8	750,000	Reversible terminators
MiSeq/Illumina/ 15 Gb	2x300	$3x10^8$	27 h	0.13	0.8	125,000	Reversible terminators
SOLiD/Life Technologies/120 Gb	50	$1x10^9$	14 days	0.13	0.01	350,000	Ligation
Retrovolocity/BGI/ 3000 Gb	50	$1x10^9$	14 days	0.01	0.01	$12x10^6$	Nanoball/ ligation
Ion PGM/Life Technologies/100 Gb	200	$6x10^7$	2-5 h	1	1.7	215,000	Proton detection
Ion Proton/Life Technologies/2 Gb	200	$5x10^6$	2-5 h	1	1.7	80,000	Proton detection

Preparation for Sequencing

While the various platforms may vary in their chemistries, most share a common set of preparation features. These steps are performed prior to sequencing and generally include (1)

amplification of the target, (2) fragmentation, (3) end repair, (4) A-tailing, (5) adapter ligation, and (6) purification steps. With mtDNA, amplifying the full mitogenome can be achieved by using two long-range (LR) primer sets that target segments approximately 8,500 bp in length (Table 2.2; Peck et al., 2018). These amplicons overlap providing full coverage of the mitochondrial genome.

Table 2.2: Long-range amplification ranges and sizes. (From Peck et al., 2018)

Amplicon	**Amplicon Range**	**Amplicon size** (bp)	**Target Range** (excludes primers)
LRA	np 2480-10858	8379	np 2500-10838
LRB	np 10653-2688	8605	np 10673-2668

Amplification occurs via polymerase chain reaction (PCR). PCR is a method of synthesizing target DNA to generate numerous copies of that segment of DNA (Hadidi and Candresse, 2003). The selected primer pairs are complementary to the 5' and 3' end of the target segment of DNA. DNA polymerase, the enzyme that synthesizes DNA, is added in combination with dNTPs necessary for creating new strands of the target DNA. When appropriate thermal conditions are applied, a series of reactions will occur. The double-stranded DNA will denature, forming single strands and the primers will anneal to their complementary sequence of target DNA. DNA polymerase will add the appropriate dNTPs to the single-stranded DNA. This is usually carried out for 25 to 40 cycles, resulting in numerous copies of the target DNA.

Following amplification, a random fragmentation step is employed. This can be done with physical methods like acoustic shearing and sonication or by enzymatic methods like the use of non-specific endonuclease mixtures and transposase tagmentation reactions where the DNA is fragmented and tagged with adapters in a single reaction (Head et al., 2014). Library

preparation kits purchased through biotechnology companies often come with the enzymes necessary to fragment the DNA and include the appropriate temperatures and lengths of time for the DNA to be incubated with the enzyme mixture. Longer incubation periods result in smaller fragment sizes whereas shorter incubation periods result in larger fragments. The desired length of the fragments will depend on the capabilities and requirements of the sequencing platform used downstream, though 150-300 base pair lengths are most commonly required for NGS platforms.

Fragmentation often leaves DNA in a non-homogenous state with over-hanging ends. End repair ensures each molecule is free of overhangs either by trimming back or filling in the ends to the ensure they are homogenous. Additionally, this step ensures the ends have 5' phosphate groups and 3' hydroxyl groups. Phosphorylation of the 5' ends makes certain that the DNA fragments are ready for ligation (New England Biosystems, 2016). Following end-repair, an A-tailing step is often employed, however some commercially-produced kits merge the end repair and A-tailing processes into one step. A-tailing, or adenylation, is a process that incorporates non-templated deoxyadenosine 5'-monophosphate (dAMP) onto the 3' end of the end-repaired fragments. This process prevents concatamer formation in downstream ligation where long continuous DNA molecules can form that have the same sequences linked in a series. Additionally, the single A-overhang can enable DNA fragments to be ligated to adaptors with complementary T- overhangs (New England Biosystems, 2016). Both the end repair and A-tailing processes require thermal incubation for optimal results.

Adapters or primer tags may include several elements: specific sequences for clonal amplification of the library, target sequences for the NGS reaction, a key sequence with 4 to 8 nucleotides used for quality control of the NGS reaction, and a 6 to10 nucleotide barcode that is

used to identify the sample (Børsting and Morling, 2015). The order of these elements is varied depending on the NGS platform used. The adaptors are ligated to the DNA fragments through a thermal incubation with DNA ligase (Mardis, 2013). Once adapters have been added to the DNA fragments they become known as "libraries."

Barcoding makes multiplexing possible. Barcodes are specifically used in the post-sequencing process to bioinformatically separate each sample from all other samples. Because each sample library will be pooled into one tube with all other sample libraries in a downstream step, adding these identifying barcodes is crucial and allows for many samples to be sequenced simultaneously. These are generally supplied by the biotechnology company of the platform chosen for use and come with a variable number of unique barcode sequences. Depending on the number of samples to be sequenced, more than one barcode can be used per sample. If the sequencing run contains only a few samples, only one barcode is added per sample. However, if many samples are to be sequenced, two barcodes may be used per sample, one at each end of the DNA fragment. Further, three or more barcodes can be used in varying combinations. This increases the number of unique barcodes as they can be combined in a number of unique ways.

An example can be seen in Figure 2.1 from Illumina's ForenSeq (2015) indices. There are 12 unique adapters of Index 1 (red) along the top of the plate, running from column 1 to column 12. There are 8 unique adapters of Index 2 (white) along the side of the plate, running from row A to row H. While the same Index 1 sequence will be found in all of column 1, each well will contain a different Index 2 sequence. Additionally, while all of row A will have the same Index 2 sequence, each well will contain a different Index 1 sequence. This demonstrates how two barcodes can be used simultaneously to uniquely tag 96 individual samples. Using more than 2 barcodes, thousands of samples can be sequenced simultaneously (Liu et al., 2012).

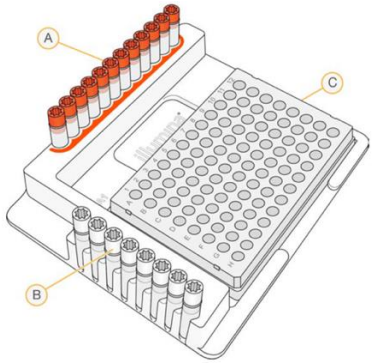


Figure 2.1: Example of Barcode Multiplexing. Twenty barcodes can be used to create 96 unique barcode combinations for a full plate. From Illumina Inc., 2015.

A purification step is often required after the steps of library preparation and after amplification to remove any remaining primers, adaptors, dimer molecules, unincorporated nucleotides, salts, and enzymes from earlier steps (Beckman Coulter, 2009; Head et al., 2013; Tan and Yiap, 2009). This can be done in several ways however magnetic or paramagnetic bead-based purification is used most often because of its efficiency and ability to be automated using robotics, which is preferred in high throughput research. Bead-based cleanup involves magnetic beads that bind to DNA (Elkin et al., 2001). The beads are made of magnetizable cellulose, magnetizable cellulose derivatives, porous glass, iron-oxide, or synthetic polymers which bind to DNA in the presence of certain concentrations of salt (Davies et al., 1997; Levison et al., 1998; Nargessi, 2005). Additionally, the particles can be coated with carboxylic acid or streptavidin, however this coating and the addition of a functional group leave less surface area on the beads for the binding of DNA (Saiyed et al., 2006).

STR Sequencing Preparation

Next-generation sequencing of autosomal and Y-STRs is often performed using commercially available kits as they contain numerous primer sets and reagents necessary for amplification at multiple loci. Additionally, there are fewer preparatory steps that take place prior to sequencing. For most kits, a primer mix that contains a pair of tagged oligonucleotides for each target sequence is mixed with each sample. PCR cycles attach the tagged oligos to the copies of each target. This forms DNA templates that contain the regions of interest that are flanked by universal primer sequences (Fig. 2.2). Indexed adapters (discussed previously) are attached to the tagged oligos and amplified again using PCR. The libraries are then purified, quantified, and normalized using normalization beads to ensure samples are equally represented within the sequencing run (Illumina, 2015). Normalization beads work by binding to DNA and eluting off the beads at approximately the same concentration for each sample (Illumina, 2017). Finally, the libraries are pooled and diluted prior to sequencing on one of the NGS platforms discussed below.

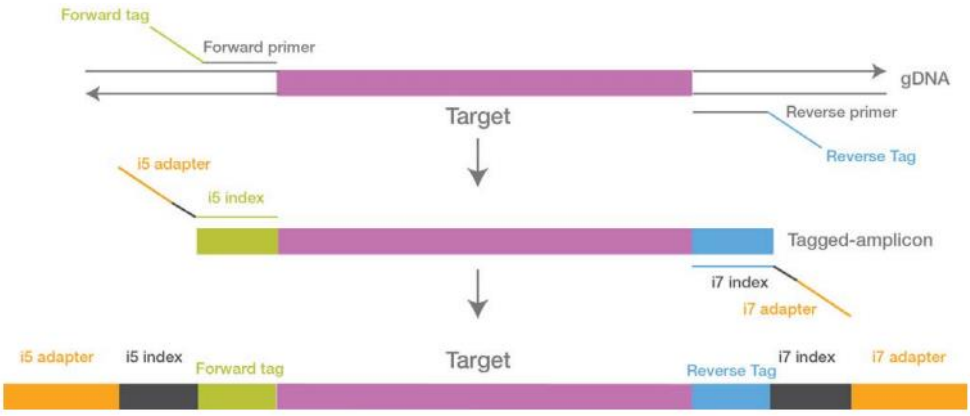


Figure 2.2: ForenSeq DNA Signature Preparation. From Illumina, 2015.

Pyrosequencing by Synthesis

While many of these NGS platforms share similar preparatory steps, they utilize a variety of sequencing chemistries (outlined in Table 2.1). The pyrosequencing by synthesis platforms (454 FLX+ Roche and GS Junior Roche) were some of the first commercially successful NGS platforms (Kulski, 2016). In this process, micro-scale beads are utilized. The aforementioned adapters attached to the target sequence are complementary to an oligonucleotide that is bound to the surface of the beads. This creates a situation whereby one DNA molecule per bead is favored Metzker, 2010). The DNA template hybridizes to the bead-bound primers. The beads are then deposited onto a PicoTiterPlate with etched microwells (Leamon et al., 2003). Amplification takes place through emulsion PCR where each individual molecule of fragmented DNA is captured on a separate bead. Each bead is separated into a droplet of PCR reaction mixture within an oil emulsion (Holt and Jones, 2008; Shendure and Ji, 2008). The DNA is amplified clonally on the surface of the bead, resulting in 100-200 million beads with thousands of bound template (Goodwin et al., 2016). The beads containing the amplified products are then immobilized in these microwells and fluorescently labeled probes are hybridized to their targets. The PictoTiterPlate is scanned at varying wavelengths, producing images of the fluorescent probes at different wavelengths ensuring amplification was successful (Leamon et al., 2003; Mardis, 2013).

Following this, pyrosequencing occurs. Nucleotides are added sequentially to the DNA synthesis reaction. Wells of the PicoTiterPlate are loaded with sequencing enzymes bound to beads: polymerase, sulfurylase, and luciferase. A buffer containing one of four dNTPs is passed horizontally over the wells. If a match to the primed template is found, the polymerase enzyme will incorporate the nucleotide and will release a pyrophosphate molecule. This molecule is

converted to ATP by the enzyme sulfurylase and generates a luminometric signal that is catalyzed by the enzyme luciferase (Garrido-Cardenas et al., 2017; Holt and Jones, 2008; Mardis, 2013). The residual nucleotides are washed away or are removed via the action of an apyrase enzyme to avoid residues interfering in later cycles (Garrido-Cardenas et al., 2017). The cycle is repeated with the next dNTP (Holt and Jones, 2008). Each burst of light is attributed to the incorporation of one or more identical dNTPs and is captured by a CCD camera (Goodwin et al., 2016).

Sequencing by Synthesis with Reversible Terminators

With sequencing by synthesis using reversible terminator chemistry (Illumina HiSeq, MiSeq), bridge amplification is first performed on a flow cell. A flow cell is an optically transparent disposable glass surface that contains a lawn of high-density forward and reverse primers. These primers are covalently bound to the surface of the flow cell and are complementary to the adapter-added DNA fragments (Nuwaysir et al., 2002; Holt and Jones, 2008; Metzker, 2010; Harakalova et al., 2011; Goodwin et al., 2016). First, the fragmented DNA is joined to a pair of oligonucleotides with a forked adaptor configuration (Fig. 2.3). This is amplified using two oligonucleotide primers that leave the template/oligonucleotide material with a different adaptor sequence on either end.

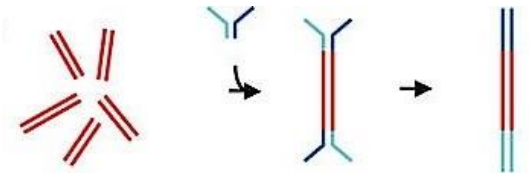


Figure 2.3: Bridge Amplification Step 1. DNA fragments are randomly sheared. Fragments are then joined to oligonucleotides with forked adaptor configuration with different adaptor sequences on each end. This product is amplified. From Bentley et al., 2008.

Once adaptors have been added to the double-stranded fragments, they are denatured. The single strands are annealed to complementary oligonucleotides on the surface of the flow cell (Fig. 2.4). A new strand is made from the original strand via DNA polymerase and the original strand is then removed via denaturation. The 3' end of the adaptor sequence of each copied strand is annealed to a new complementary oligonucleotide that is bound to the surface of the flow cell. This forms a 'bridge' and generates a new site for synthesis of a second strand (Bentley et al., 2008).

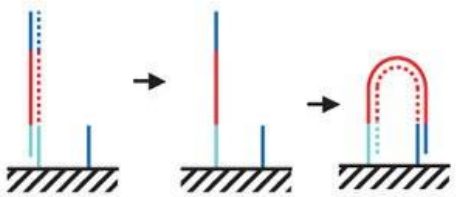


Figure 2.4: Bridge Amplification Step 2. Double-stranded fragments are denatured to single-stranded fragments to be attached to complementary oligonucleotides on the flow cell. A new strand is synthesized and then denatured. The 3' end of the adapter sequence is annealed to a new complementary oligonucleotide that is bound to the flow cell, forming a bridge. Isothermal bridging amplification follows. From Bentley et al., 2008.

Isothermal conditions of multiple cycles of annealing, extension, and denaturation result in the growth of DNA clusters which are approximately 1 µm in diameter (Fig. 2.5 and 2.6). The formation of these clusters is important in downstream sequencing as signals from individual strands are often insufficient for detection, which is why bridge amplification and the subsequent formation of DNA clusters is necessary.

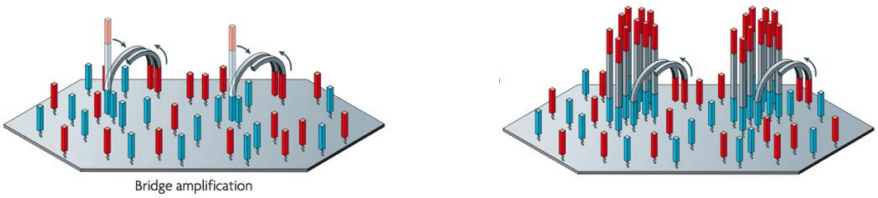


Figure 2.5: Bridge Amplification Step 3. Amplification on a flow cell. Numerous cycles of bridge amplification result in clusters of amplified DNA attached on a flow cell. From Metzker, 2010.

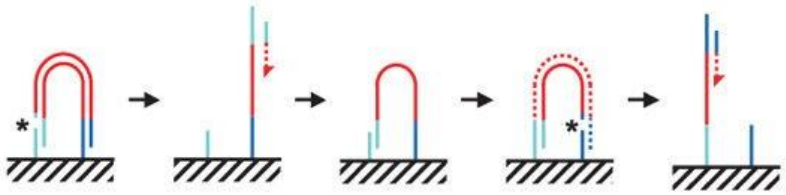


Figure 2.6: Bridge Amplification Step 4. The DNA in each cluster undergoes cleavage of one adaptor sequence and denaturation, generating a single-stranded template used for sequencing by synthesis. The products of the first read are denatured and the template is used to create a bridge. The second strand is re-synthesized while the opposite strand is cleaved. This leaves a template for the second read. From Bentley et al., 2008.

To sequence the ends of long DNA fragments, those longer than 1 kb, the ends of each fragment are tagged by incorporation of a biotinylated nucleotide (Bentley et al., 2008; Mardis, 2013). They are then circularized, forming a junction between the two ends (Fig. 2.7). Instead of the normal two adapters, only one adaptor is used. Referred to as mate-pair sequencing, the circularized DNA is again randomly fragmented and the biotinylated junction fragments are used as the beginning material in the standard sample preparation procedure shown in Figure 2.3. The biotinylated adapter can be captured using streptavidin magnetic beads during the washing process. The process is then repeated where adapters are ligated, clusters are generated, and the first end is sequenced followed by the sequencing of the paired end (Mardis, 2013).

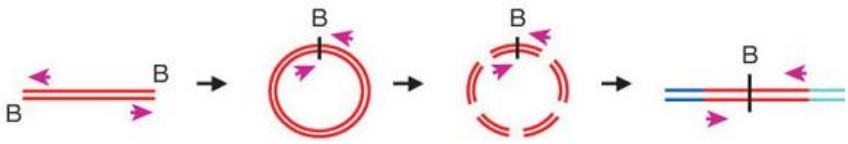


Figure 2.7: Circularized DNA fragmentation. A biotinylated nucleotide is added to the ends of each fragment (Denoted "B"). The two ends of the DNA fragments are joined, creating circularized DNA. The circularized DNA is randomly fragmented. The biotinylated junction fragments are used as the beginning material seen in Figure 2.3. From Bentley et al., 2008.

Following bridge amplification, the DNA is linearized in each cluster of strands by cleaving one adaptor sequence and denaturing it. This generates single-stranded template that is

used for sequencing by synthesis (Fig. 2.3) (Bentley et al., 2008). The instrument sequences from one adapter priming site using a stepwise sequencing strategy. The products of the first read are removed by denaturation. This is followed by a subsequent reaction where the template is used to create a bridge while the second strand is re-synthesized and the opposite strand is then cleaved, leaving the template for the second read.

This sequencing chemistry is similar to Sanger sequencing but they differ in that the obstruction of DNA polymerization is reversible whereas in Sanger sequencing, it is irreversible (Garrido-Cardenas et al., 2017). A nucleotide is first added in each cluster on the surface of the flow cell by incorporating one of the four nucleotides of reversible termination. The nucleotides that are incorporated are chemically blocked by the substitution of the 3'-OH group for a 3'-*o*-azidomethyl group (Bentley et al., 2008). This prevents the polymerase from incorporating more than the one nucleotide in each cycle (Garrido-Cardenas et al., 2017). The nucleotide added is detected by various laser channels and measured by the total internal reflection fluorescence. Any unincorporated nucleotides are then washed away and the 3' blocking groups are removed through the application of tris-(2-carboxyethyl)-phosphine. This step is performed to continue the synthesis of the chain, making them cyclic reversible terminators. Rather than being distinct processes, the platforms sequence and detect simultaneously (Mardis, 2013).

Sequencing by Ligation

With sequencing by ligation platforms (Life Technologies SOLiD and BGI Retrovolocity), DNA ligase is used instead of DNA polymerase (Shendure and Ji, 2008). With SOLiD, following fragmentation, two adaptors are attached to the ends of the fragments, termed P1 and P2 (Mardis, 2013). PCR follows via hybridization of beads that contain fragments corresponding to the P1 adaptor. PCR creates polonies, or polymerase colonies, on the beads

(Mitra et al., 2003). However, more PCR beads are used than there are fragments. Therefore, most of the PCR beads go unused and must be removed to conserve space on the array to be sequenced. These excess beads are removed by adding polystyrene beads that are coated with P2 adaptors. The beads that contain bound DNA fragments will have P2 adaptors at their ends. These P2 adaptors bind to the complementary P2 adaptors on the polystyrene beads. These structures will be saved for sequencing. Centrifugation removes unused beads as the PCR beads have more mass and will move to the bottom while the supernatant will contain the unused beads (Applied Biosystems, 2011).

The polonies are bound covalently to a glass slide. An 8-nucleotide probe with an attached fluorophore is used. As with most sequencing chemistries, the fluorophore corresponds to one of the four dinucleotide possibilities (Applied Biosystems, 2011; Liu et al., 2012). Of this 8-base probe, the first two nucleotides are actual bases, one of the 16 possible permutations. The following 3 bases are universal bases than can bind to any of the four nucleotides. The last three nucleotides of the probe are universal bases with fluorescent dye. Once these are measured for fluorescence, they are then cleaved in each cycle so that the attached probe is only 5 nucleotides in length (Garrido-Cardenas et al., 2017).

A primer is added that corresponds to the P1 adaptor and anneals to the fragment to be sequenced. A probe is attached right after the primer via DNA ligase. Additional probes are ligated directly after subsequent probes, such that probes are ligated to one another sequentially. The fluorescence snapshot takes place where a laser excites the fluorescent dye. It then releases a lower energy photon which is detected and recorded. The dye end is then cleaved, leaving 5 nucleotides on the probe and a free 5'-phosphate end (Liu et al., 2012). This allows for the next probe to attach and be ligated. As seen in Figure 2.8, the first and second base of each ligation

reaction is learned, occurring every 5 bases (Applied Biosystems, 2011). This process is repeated with the exception that each cycle is offset by one base. Each base position is queried twice, at the first base and the second base, in a set of 2 base pair examination in a cycle (Shendure and Ji, 2008). This process is repeated for five rounds of primer reset (Applied Biosystems, 2011). When the results of all cycles are overlaid, the entire sequence can be learned.

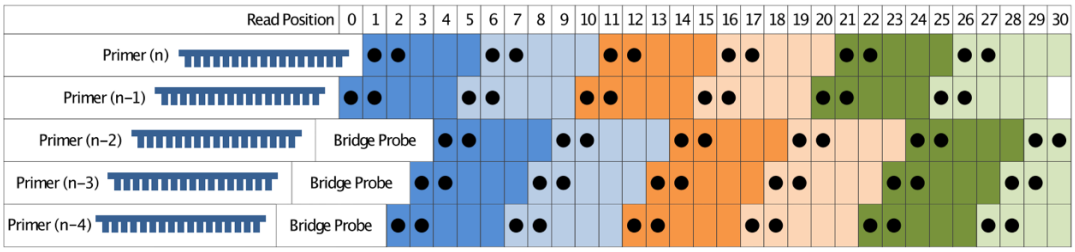


Figure 2.8: SOLiD Sequencing by Ligation. Two-base probes are interrogated and five cycles of offset probes reveal the target sequence. From Applied Biosystems, 2011.

The SOLiD sequencing technology does not work well with palindromic sequences. Palindromic sequences read the same no matter which 5' end of the sequence is read. These present issues as SOLiD technology requires single-stranded templates for the oligonucleotides to hybridize. The palindromic sequences have a high affinity for self-annealing when in a single-stranded form, creating hairpins within the single-stranded template. This makes the DNA inaccessible to the probes and unable to be sequenced (Huang et al., 2012). Despite this, sequencing by ligation produces the lowest error rate among many of the NGS methods (Børsting and Morling, 2015).

The BGI Retrovolovity works slightly different from the SOLiD technology. Following fragmentation, the DNA is ligated to the first of four adapter sequences (Goodwin et al., 2016). The template is amplified, circularized, and then cleaved with an endonuclease (Fig. 2.9). A

second adapter set is added, amplified, circularized, and cleaved. This same process is repeated for the remaining two adapter sets. This results in a circular template with four adapters separated by a template sequence. Following this, the molecules are amplified by rolling-circle amplification. This creates a mass of concatamers known as nanoballs that are held together through intramolecular interactions (Goodwin et al., 2016). Up to 20 billion DNA nanoballs are generated during this process.

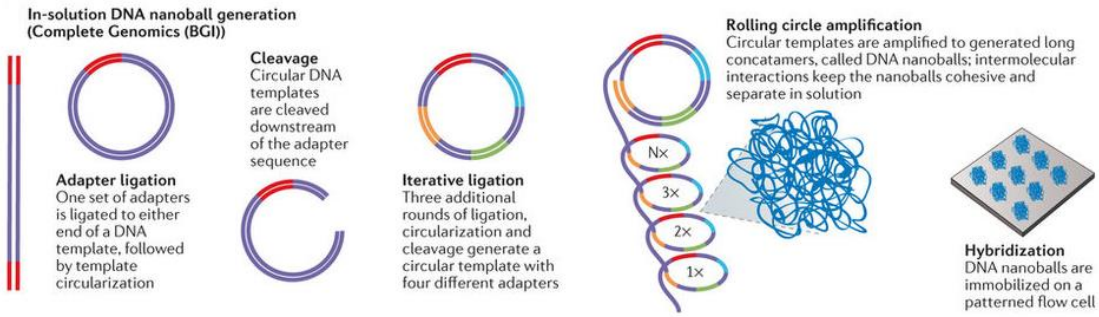


Figure 2.9: Nanoball Generation with Rolling-circle Amplification. From Goodwin et al., 2016.

The DNA nanoballs are deposited onto an arrayed flow cell with one nanoball per well (Kulski, 2016). Up to 10 bases are read in both the 3' and 5' directions from each adapter, similar to SOLiD. An anchor that is complementary to one of the four adapter sequences with an accompanying fluorophore-labelled probe are bound to each nanoball (Goodwin et al., 2016). The anchor and probe are ligated into position and imaged. This will identify the first base on both the 3' and 5' side of the anchor. The probe-anchor complex is then removed and the process is repeated with the same anchor but different probe with the identified base at the n+1 position (Fig. 2.10). This is repeated until five bases from the 5' and 3' ends of the anchor are identified. Following this, the ligated sequencing probes are removed and a new pool of probes is added. The whole process is repeated for each of the remaining three adapter sequences in the nanoball, generating 100bp paired-end reads (Goodwin et al., 2016; Kulski, 2016).

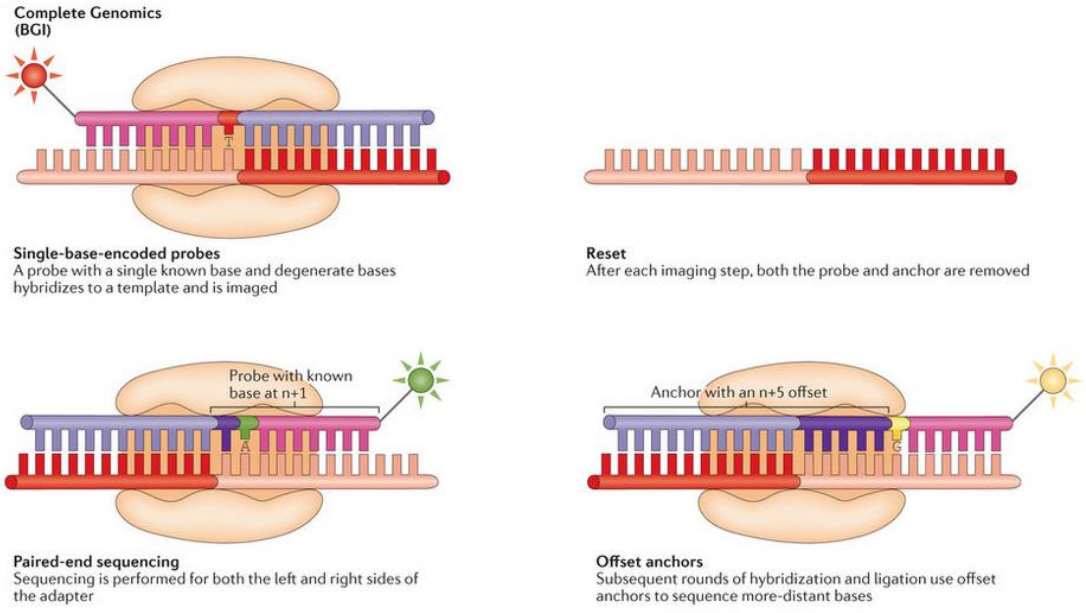


Figure 2.10: Complete Genomics/BGI Retrovolocity Sequencing. From Goodwin et al., 2016.

Sequencing by Proton Detection

Unlike the previous platforms, proton detection platforms (Life Technologies Ion PMG; Ion Proton) do not use optical sensing but instead monitor pH changes. This approach not only is one of the only methods not utilizing optical sensors but also eliminates the use of dNTPs attached to fluorophores (Garrido-Cardenas et al., 2017). Nucleotide incorporation produces the release of hydrogen ions (Mardis, 2013). The release of these hydrogen ions can be determined through the detection of changes in pH. The changes in pH are detected by an integrated complementary metal-oxide-semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) (Garrido-Cardenas et al., 2017; Goodwin et al., 2016). Following library construction and bead-based emulsion PCR (detailed in the Pyrosequencing by synthesis section), the beads are primed for sequencing by annealing a sequencing primer. They are then deposited into wells of a chip, such as an Ion Chip produced by Ion Torrent. The addition of a single nucleotide

occurs one at a time so it is not necessary to block the dNTPs like with cyclic reversible terminator sequencing (Garrido-Cardenas et al., 2017).

The specialized silicon chip is specifically designed to detect changes in pH within each individual well as the stepwise reaction progresses (Mardis, 2013). In the case of the Ion Chip, one surface serves as a microfluidic conduit that delivers the reactants necessary for the sequencing reaction while the other surface interfaces directly with the hydrogen ion detector (Fig. 2.11). It translates the released hydrogen ions from each of the wells into details of the nucleotide bases that were incorporated in each reaction step. The detection of pH changes can prove difficult though. The sensor is imperfectly proportional to the number of nucleotides detected. This means the accuracy can be flawed, particularly in homopolymeric regions (Goodwin et al., 2016).

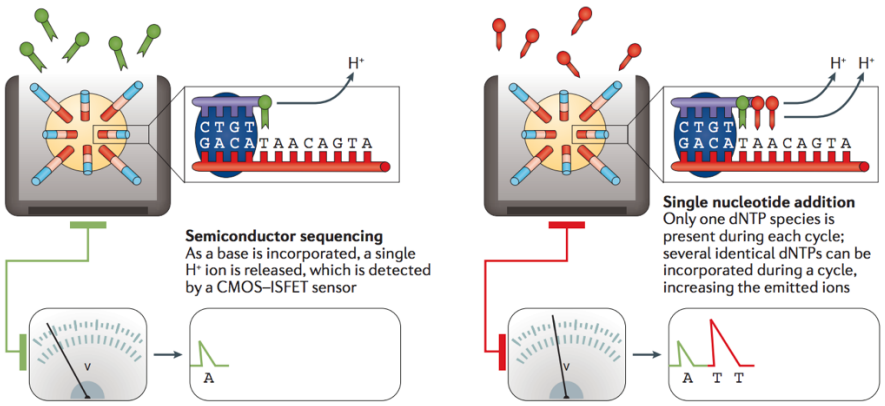


Figure 2.11: Sequencing by Detection of pH change. Each well of the chip acts as a pH meter. From Goodwin et al., 2016.

NGS Software

With higher throughput sequencing, a need arose for accompanying software systems to analyze and assemble sequence data. Specific software often accompanies the different NGS

platforms. These software systems match or align a large number of reads against a human genome reference, usually the revised Cambridge Reference Sequence (Andrews et al., 1999). The alignment takes place in two steps. First, a limited number of candidate positions are identified using fast heuristic approaches. Instead of reading the entirety of each strand at once, only 20-40 bases of each strand are used to identify where in the genome it belongs (Berglund et al., 2011). This can be thought of as an initial sorting of the DNA. Next, the candidate positions are evaluated by methods that are more accurate, like the Smith-Waterman algorithm (Li and Homer, 2010). Essentially, a balance must be achieved between the speed and the sensitivity of these algorithms (Berglund et al., 2011). Generally, the faster the alignment algorithm, the less accurate it will be.

An evaluation of Sanger sequencing versus NGS found that sequence alignment algorithms used to align NGS reads played a significant role in the data analysis (Parson et al., 2013) and thus, the resulting sequence data. This is mainly due to the inability to manually analyze NGS data or manually call genotypes (Alvarez-Cubero et al., 2017). This means that the accompanying NGS software programs need to be reliable and accurate. These programs not only assemble sequence reads but analyze the data associated with each sequence read.

Because high-throughput sequencing technologies typically produce higher error rates and genotypic uncertainty due to random chromosome sequencing (Maruki and Lynch, 2015), their error rates must be considered when determining the authenticity of the sequence data. The Illumina platform has sequencing error rate of 10^{-2} to 10^{-3} (1 nucleotide in 100-1,000 bases) (Kircher and Kelso, 2010; Fox et al., 2014; Sharma et al., 2017). The most common errors are single nucleotide substitutions resulting from errors during amplification and sequencing due to polymerase mistakes and incorrect base calling by the analysis software. To address this, a

quality score (Q score) is calculated for each run after the 25th cycle (Sharma et al., 2017). The Q score reflects the statistical likelihood that the called base is correct (Green, 2001). A Q30 score (Q30 = error probability of 0.001) and above is considered of high quality (Sharma et al., 2017). The quality score allows for the detection of authentic overlaps among sequence reads more accurately. This is in part determined by the addition of control DNA. Generally, each flow cell is sequenced with the phage phiX. The phiX reads are identified by a comparison to the phiX genome. This is used to determine the error rate of the run as well as a general measure of quality of the run overall (Berglund et al., 2011).

Many NGS systems identify clusters within the first four cycles of sequencing. If the initial recognition of these clusters is imperfect, it may indicate that clusters contain more than one originating template. Additionally, this can be a sign of phasing or pre-phasing. This is where some clusters contain molecules that have incorporated fewer or more nucleotides than the number of cycles (Berglund et al., 2011). These clusters can be filtered out based on signal intensity ratios of all clusters on the flow cell. NGS software programs also include information about the sequencing depth or coverage, meaning the number of reads that sequenced that exact base position. A variable degree of misidentifications or misincorporations occur at a low rate thus high coverage is needed to ensure base calling is correct (Rizzo and Buck, 2012).

Chapter 3: NGS using MtDNA

Traditional Methods

Mitochondrial DNA is a circular molecule 16,569 bps in length. MtDNA is inherited without recombination from mother to offspring and is informative in anthropological genetics for illuminating relationships between groups and individuals from a maternal perspective. MtDNA haplogroups are defined by a collection of shared, inherited polymorphisms. Many of these polymorphisms are located within the control region, particularly hypervariable segments 1 and 2 (Fig. 3.1), as these are known to be highly polymorphic and non-coding (Rubicz et al., 2006). This means natural selection does not act to reduce genetic variability in this region, making it useful in anthropological studies. Coding regions lie outside the control region. While the coding region traditionally is less informative for anthropological studies, there are positions that contain variants useful for population characterization and haplogrouping. With the addition of coding region polymorphisms in mitogenome NGS, precise, fully-derived haplogroups can be assigned. When only data from the HVs or CR are used, haplogroup estimations are made using the available data.

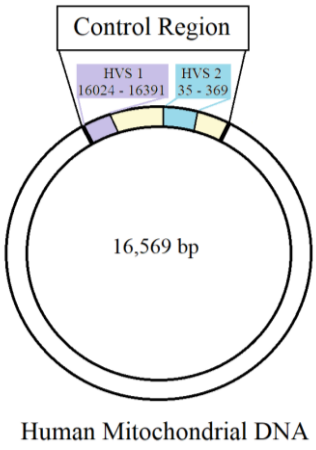


Figure 3.1: Human mtDNA with HV1, HV2, and the CR Highlighted.

mtDNA Haplogroup B

MtDNA haplogroups are used to obtain broad ancestral information in anthropological genetics. Tracing the branches of the mtDNA tree, it is possible to identify lineages that are seen in particular geographic regions that are absent elsewhere. At the root of the tree is the African macrohaplogroup L (Chen et al., 1995; Watson et al., 1997; Wallace et al., 1999; Behar et al., 2008; Cerezo et al., 2016) with successive branching of Eurasian macrohaplogroups M, N, and R representing human migration out of Africa (Fig. 3.2) (Maca-Meyer et al., 2001; Forster, 2004; Quintana-Murci et al., 2004; Macaulay et al., 2005; Soares et al., 2012).

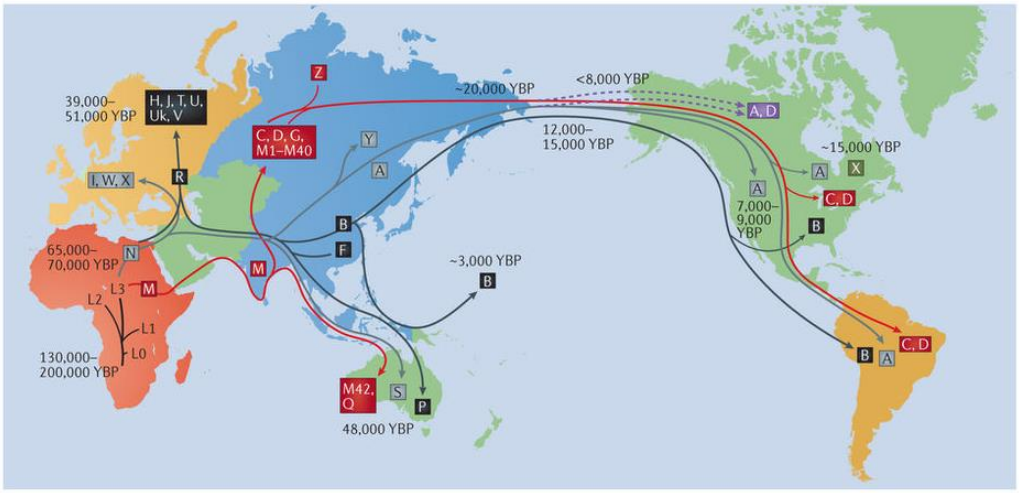


Figure 3.2: Global Spread of mtDNA Haplogroups (Stewart and Chinnery, 2015)

Within Asia, these macrohaplogroups differentiated to form numerous haplogroups that spread throughout the region. Indigenous peoples of the Americas carry five major haplogroups (A, B, C, D, X) (Torroni et al., 1993a,b; Crawford, 1998; Bandelt et al., 2003; Tamm et al., 2007; Fagundes et al., 2008; Perego et al., 2009; O'Rourke and Raff, 2010). These haplogroups are also found in Asia, suggesting settlement of the Americas occurred from migration of individuals

from continental Asia. Three of the five haplogroups (A, C, D) were also observed in Siberian-Eskimo populations (Torroni et al., 1993a; Starikovskaya et al., 1998), further suggesting this migration occurred via Beringia. Haplogroup X, specifically X2, was observed in southwestern Siberia and the Altai (Reidla et al., 2003; Derenko et al., 2001; Phillips-Krawczak et al., 2006). However, the lineages observed were later identified as non-ancestral to the Native American X2a lineages (Raff and Bolnick, 2015). The inexact linking of these two haplogroups was realized when complete mitogenome sequences were analyzed. This reinforces the importance of mitogenome sequencing in anthropology for interpreting genetic evidence of migratory patterns and understanding American prehistory. While the Siberian X2 lineages are not ancestral to the Native American X2a lineages, it is believed they still arrived in the Americas via Beringia and the intermediate lineages have since been lost or are very rare among modern populations (Raff and Bolnick, 2015).

Haplogroup B was not observed in eastern Siberia until more in-depth sequencing took place where it was reported in low frequencies in the Altai, northeastern Siberia, and most of northern Asia (Shields et al., 1992, Torroni et al., 1993b, Sukernik et al., 1996; Phillips-Krawczak et al., 2006). However, haplogroup B is seen in higher frequencies in southeastern Siberia and the Mongolia/Machuria region (Kolman et al., 1996; Merriwether et al., 1996; Derenko et al., 2003), and even higher frequencies in East Asia (Torroni et al., 1993a; Derenko et al., 2003; Tanaka et al., 2004; Starikovskaya et al., 2005; Derenko et al., 2012), suggesting this may be the originating area of this haplogroup before migration up into Beringia and the Americas.

Migration into the Americas took place relatively recently in human history (Goebel et al., 2008; O'Rourke and Raff, 2010; Raghavan et al., 2015), owing to the genetic relationship

between indigenous peoples of Asia and the Americas. Collectively, coalescence dates for Native American haplogroups fall between 17,000-34,000 years BP (Torroni et al., 1993b). Within haplogroup B specifically, the coalescent age estimation of haplogroup B2 has been calculated to 20.8 +/- 2.0 kya (Kumar et al., 2011) using a mutation rate of one base substitution in the coding region per 5,140 years described by Mishmar and colleagues (2003). A second coalescent age estimation was calculated to 18.1 +/- 2.4 kya (Kumar et al., 2011) when using mutation rates described by Soares and colleagues (2009).

Approximately 20,000 years has passed since individuals carrying haplogroup B migrated into the Americas. Over that time, mutations have arisen and have been inherited by subsequent generations. Mutations that occurred in the time since the Siberian/Native American split have become defining polymorphisms for new haplogroup subtypes. In modern populations, haplogroup B4 and its subtypes (excluding B2) are found in central and southeastern Asia, primarily and along the coast (Torroni et al., 1993a; Tanaka et al., 2004; Starikovskaya et al., 2005; Derenko et al., 2012). Haplogroup B2, (a subgroup of B4b) is one of few haplogroups found exclusively among indigenous peoples of the Americas (Achilli et al., 2008; O'Rourke and Raff, 2010; Achilli et al., 2013). Individuals containing an early B4b haplogroup likely migrated into Beringia and over time, accrued the B2 defining mutations that all Native American haplogroup B individual share today (Fig. 3.3: 3547G, 4977C, 6473T, 9950C, 11177T). Haplogroup B4b1 (found in Asia) also diverged from B4b, accruing separate polymorphisms not seen in haplogroup B2 (Fig. 3.3).

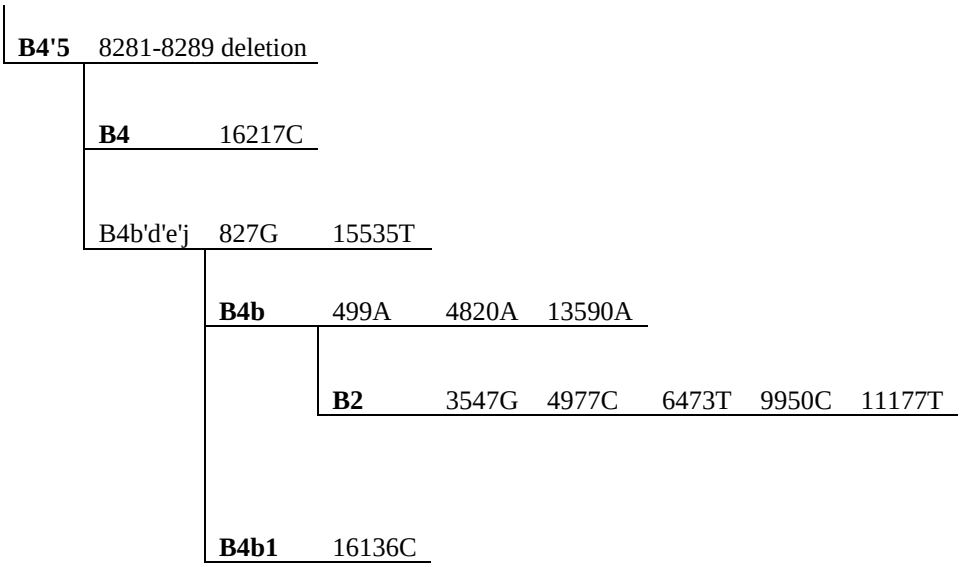


Figure 3.3: Haplogroup B4 Tree with B4b1 and B2 Branches. Adapted from Phylotree v17 (van Oven, 2015).

mtDNA Haplogrouping Methods

Native American haplogroups were the first haplogroups to be identified using restriction enzymes (Wallace et al., 1985). Particular enzymes would cut the DNA at specific sites if a particular motif was present. Haplogroup B was originally identified by the 9-bp COII-tRNA ^{Lys} intergenic deletion and an *HaeIII* site at np 16517 (Torroni et al., 1994). Today, sequencing of nucleotides using either Sanger sequencing or NGS is used to identify haplogroups. While Sanger sequencing of the HVs and CR generally provides sufficient information for general haplogrouping, this is not always sufficient.

Such an issue exists within haplogroup B. As discussed earlier, haplogroup B4b has two subclades: B4b1, found in Asia, and B2, found in the Americas. Distinguishing between haplogroups B4b1 and B2 can be difficult because the defining B2 variants are found in the mtDNA coding region, and therefore are not detected if only the smaller HV segments of the

mtDNA are interrogated. This can be problematic, particularly when trying to predict the maternal ancestral geographic origin for an individual.

Today, haplogrouping is generally performed by searching a profile in one of the available mtDNA database and haplogrouping systems such as Empop (Parson and Dür, 2007). Empop software is based on quasi-median network analysis that uses an algorithm called EMMA to perform haplogrouping. This algorithm is used for estimating the haplogroup of mtDNA sequences based on 14,990 full mitogenomes from GenBank and 3,925 virtual haplotypes from Phylotree (van Oven, 2015). Further, 19,171 control region haplotypes are used to perform a maximum likelihood estimation of the stability of mutations, expressed as fluctuation rates (Röck et al., 2013; https://empop.online/methods). When full mitogenome data are used for haplogrouping with EMMA, a more precise haplogroup can be assigned with higher confidence. When smaller regions, such as the HVs or CR are used, haplogrouping estimations are generally made with lower confidence because many haplogroups have several polymorphisms found outside the CR, though these estimations are generally conservative to avoid haplogroup misidentification.

NGS not only allows for higher confidence with haplogroup calls but can also illuminate variability within haplogroups not previously explored. With the reporting of full mitogenome variants in EMMA, private polymorphisms are often identified that are not haplogroup defining. These can be useful for tracing familial lineages and identifying levels of haplogroup variability. Further, the high coverage of most NGS platforms allows for the identification of heteroplasmies. While Sanger sequencing is also capable of detecting heteroplasmies, NGS can detect these at a much more sensitive level, opening up new areas for future research including

calculating prevalence rates of mtDNA heteroplasmies and the identification of rate differences between varying tissue types.

Chapter 4: Mitochondrial DNA Analyses

Materials & Methods

Samples

Anonymized DNA samples and extracts used for this research are housed at the Armed Forces Medical Examiner System's Armed Forces DNA Identification Laboratory (AFMES-AFDIL) in Dover, Delaware. Chosen samples belong to a subset of a large collection (Appendix A) that were previously sequenced for mtDNA control region data using Sanger sequencing (published in Irwin et al., 2007) and were haplogrouped by EMMA (Röck et al., 2013). These samples were not previously typed for any STR information. The use of these samples was reviewed by the U.S. Army Medical Research and Materiel Command's Office of Research Protections, Institutional Review Board Office and determined not to involve human subjects research (IRBO # M-10185).

Fifty-six samples from this collection were chosen for analysis on the basis that they reported a B2 or B4 haplogroup using previous control region data. Of these, 28 samples are self-reported Native American or Hispanic individuals while the remaining 28 samples are self-reported Asian individuals (Outlined in Table 4.1). Self-identified Asian samples are from China (DNA Diagnostics Center), Japan (DNA Diagnostics Center), the Philippines (DNA Diagnostics Center), the Altai Mountains in Siberia (Michael Crawford), Siberian Yakut individuals (Michael Crawford and Larissa Nichols), and Asian Americans (Dept. of Defense Serum Repository). Self-identified Native American and Hispanic samples are from Illinois (Dept. of Defense Serum Repository), Ohio (DNA Diagnostic Center), South Dakota (South Dakota Dept. of Public Safety), Texas (Dept. of Defense Serum Repository), and Washington (Dept. of Defense Serum Repository). Appropriate controls were processed with all samples including a negative control

(NC), positive control (PC) (2800 M, Promega, Madison, WI), and reagent blanks (RBs) from previously performed DNA extractions.

Table 4.1: List of Samples used for mtDNA Analyses.

mtDNA Samples	
Affiliation	**No. of Samples**
Chinese	5
Japanese	6
Filipino	9
Asian American	2
Siberian	6
Native American: South Dakota	6
Native American: Washington	6
Hispanic: Illinois	8
Hispanic: Texas	4
Hispanic: Ohio	4
Total:	56

Amplification

PCR amplification of 60 total samples was performed: 56 samples, one PC, one NC, and two RBs. Two long-range primer sets were used from a previously published method (Gonder et al., 2007) for targeting two long-range amplicons that overlap to capture the full mitogenome (Table 4.2). Amplification reactions included 0.25 µL of Advantage GC Genomic LA Polymerase Mix (Clontech, Mountain View, CA), 12.5 µL of 2X Advantage Genomic LA Buffer (Clontech), 4 µL of 2.5 mM dNTPs (Applied Biosystems, foster City, CA), 1.25 µL of nuclease-free water, 1 µL each of 10µM primers (Integrated DNA Technologies Inc., Coralville, Iowa), and 5 µL of DNA template, for a total of 25 µL reactions. Thermal cycling was performed using a Veriti Thermal Cycler (Applied Biosystems) adhering to the conditions outlined in Table 4.4.

Table 4.2: Amplicons and Primers used for Long Range Amplification of mtDNA. Previously published in Gonder et al., (2007).

Amplicon	Target Range	Amplicon size	Primer	Sequence 5'-3'
LRA	np 2500-10838	8379 bps	For 2480	AAATCTTACCCCGCCTGTTT
			Rev 10858	AATTAGGCTGTGGGTGGTTG
LRB	np 10673-2668	8605 bps	For 10653	GCCATACTAGTCTTTGCCGC
			Rev 2688	GGCAGGTCAATTTCACTGGT

When samples failed to amplify, a second primer set was attempted for LR amplification using previously published primer sets (Table 4.3; Tanaka et al., 1996) and using the thermocycler conditions outlined in Table 4.4. These primer sets successfully amplified three additional samples.

Table 4.3 Amplicons and Primers used for Long Range Amplification when First Primer Set Failed. Previously published in Tanaka et al. (1996).

Amplicon	Target Range	Amplicon size	Primer	Sequence 5'-3'
LRA	np 2835-11548	8754 bps	For 2817	GCGACCTCGGAGCAGAAC
			Rev 11570	GTAGGCAGATGGAGCTTGTTAT
LRB	np 10816-3350	9144 bps	For 10796	CCACTGACATGACTTTCCAA
			Rev 3370	AGAATTTTTCGTTCGGTAAG

Table 4.4: Thermocycler Conditions for Long Range Amplification of mtDNA.

Thermocycler Conditions		
93°	3 minutes	
93°	15 seconds	
60°	30 seconds	14 cycles
68°	5 minutes	
93°	15 seconds	
55°	30 seconds	27 cycles
68°	9 minutes	
4°	∞	

Amplified products were quantified on a Fragment Analyzer (Advanced Analytical, Ames, IA) using the dsDNA 75-15,000 bp reagent kit (Advanced Analytical). Based on the

quantification data, both amplicons were pooled by concentration. Each amplicon pool was brought to a 25 µL working volume using 10 mM Tris-HCl (pH 8.5). Diluted amplicon pools were purified using a 1X AMPure XP (Beckman Coulter, Indianapolis, IN) bead-based cleanup. The purified amplicon pools were quantified using the dsDNA 75-15,000 bp reagent kit (Advanced Analytical) and Fragment Analyzer to assess accurate amplicon concentration for library preparation input.

Library Preparation & Sequencing

Libraries were prepared using the KAPA HyperPlus PCR-Free protocol (KAPA Biosystems, Wilmington, MA) in half-volume reactions following the protocol described in Ring et al. 2017. A total volume of 17.5 µL of template was input into the reaction and NEXTflex-HT Barcodes (Bioo Scientific, Austin, TX) were used for sample indexing. Quantitative PCR (qPCR) of purified constructed libraries was performed using the KAPA SYBR FAST qPCR Kit (Kapa Biosystems, Wilmington, MA) on an ABI 7500 Real-Time PCR instrument (Thermo Fisher Scientific, Waltham, MA) following the manufacturer's recommended protocol (Applied Biosystems, 2010).

Quantified libraries were normalized to 2 nM and then pooled in equal volume. The library pool was denatured and diluted to 10 pM according to the Illumina protocol for MiSeq sequencing (Illumina, 2016). PhiX Sequencing Control v3 (Illumina) was spiked into the pool at 2.5% prior to loading. Single-end sequencing was performed using a 150-cycle MiSeq v3 Reagent Kit (Illumina, San Diego, CA) on a MiSeq FGx Instrument (Illumina) in Research Use Only mode.

Data Analyses

MiSeq Reporter (Illumina) generated FASTQ files and the data were analyzed using the CLC Genomics Workbench (version 7.5.1) and the AFDIL-QIAGEN mtDNA Expert tools (AQME; Sturk-Andreaggi et al., 2017). The analysis workflow included: 1) 20 bp were trimmed on the 5' and 3' ends of all sequencing reads, 2) reads were mapped to the revised Cambridge Reference Sequence (rCS; Anderson et al., 1981; Andrews et al., 1999), 4) the mapping was locally realigned to improve indel alignment, 4) quality-based (≥30) variant detection requiring 10X coverage and 5% variant frequency (VF) with 5% forward-reverse balance, and 5) forensic profile generation with AQME (Sturk-Andreaggi et al., 2017). A read direction filter of 1% was used to confirm the HV 2 polycytosine stretch (i.e. presence of 315.1C).

Haplogrouping was performed using increasing ranges of mitochondrial DNA data: HV1 (16024-16391), HV1 &2 (16024-369), CR (16024-574), and full mitogenome (1-16569). Variants present in each region were input separately into the haplogrouping system SAM2 in EMPOP4 (W. Parson, personal communication; Huber and Parson., in prep). Haplogroup assignments were produced from each of the 56 samples for each of the range definitions. To examine the ability to discern Native American haplogroups from Asian haplogroups, maximum parsimony phylogenetic trees were constructed using MtPhyl (http://eltsov.org). Each tree was adjusted manually according to Phylotree v17 (van Oven, 2015).

Results

Haplogrouping Accuracy

The number of distinct haplogroups identified in the data and the number of samples belonging to each haplogroup are listed in Table 4.5. Of the 56 samples, 28 were of Native American origin and 28 were of Asian origin. Ten of the Native American samples belonged precisely to haplogroup B2 while 18 belonged to B2 subgroups. As expected, all Asian samples belonged to B4 subgroups. Mitogenome sequencing confirmed there were no overlapping haplogroups between the Asian and Native American samples. None of the Asian samples belonged to haplogroup B2 or B2 subgroups and none of the Native American samples belonged to B4 subgroups. Among the 56 samples, there are 53 unique haplotypes represented.

Table 4.5: Number of Haplogroups Present.

Haplogroup	No. of samples		Haplogroup	No. of samples
B2	9		B4a4	5
B2+16278	1		B4a1a5	1
B2a	1		B4b1	1
B2a1	2		B4b1a1	1
B2b	1		B4b1a1a	1
B2c	3		B4b1a2	4
B2c2a	2		B4b1a2b1	1
B2f	1		B4b1a3a	3
B2g1	1		B4b1b	1
B2k	1		B4c1a1a	1
B2l	1		B4c1b1a	1
B2o	1		B4c1b2a2	5
B2q	1		B4c1c	1
B2s	1		B4c1c1	1
B2t	1		B4g1a	1
B2x	1		Total	28
Total	28			

Table 4.6 outlines the haplogrouping results for each sample using four ranges of mtDNA. Of the 56 samples, 50% (28/56) could not be precisely haplogrouped without sequencing the full mitogenome, meaning the precise haplogroup (fully derived haplogroup) could not be identified without information found outside the CR. Using only HV1 data, precise haplogroups were determined for 32% (18/56) of the samples. An additional 11% (6/56) of the samples could be precisely haplogrouped when data from HV 1& 2 were included while CR data produced an additional 7% (4/56).

Bolded samples in the table represent inaccurate haplogrouping. In 5.3% of samples (3/56), inaccurate haplogroups were assigned when less than the full mitogenome was used (SDNA126, WANA007, WANA050). Each of these samples contains private polymorphisms within the CR or HVs that are also found in other haplogroups (polymorphisms are listed in Table 4.7), however each sample lacked the additional variants found in the coding region necessary for that haplogroup assignment. This led to inaccurate haplogrouping in the absence of haplogroup-diagnostic coding region variants. Highlighted samples in the table represent samples where data from the full mitogenome was needed to distinguish haplogroup B2 from B4b for discerning Native American haplogroups from Asian haplogroups.

Table 4.6: Haplogrouping Results using Data from Four Ranges. Bolded samples represent inaccurate Haplogrouping when less than the mitogenome is used. Highlighted samples represent samples where the mitogenome was needed to distinguish B2 from B4b.

Sample	Population	HV1	HV1&2	CR	Mito genome
CHN094	Chinese	B4g	B4g	B4g	B4g1a
ILH012	Illinois Hispanic	B4	B4	B4b	B2
ILH017	Illinois Hispanic	B2a	B2a	B2a	B2a1
ILH070	Illinois Hispanic	B4	B4	B4b	B2b
ILH071	Illinois Hispanic	B4	B4	B2	B2c2a

ILH087	Illinois Hispanic	B2	B2	B2	B2+16278
ILH097	Illinois Hispanic	B4	B4	B2	B2
JPN080	Japanese	B4b1a1	B4b1a1	B4b1a1	B4b1a1a
JPN260	Japanese	B4c1b1	B4c1b1	B4c1b1	B4c1b1a
OHHis068	Ohio Hispanic	B2+16278	B2+16278	B2+16278	B2k
PHL012	Filipino	B4b1	B4b1a+207	B4bla+207	B4b1a2
PHL109	Filipino	B4b1	B4b1a+207	B4b1a+207	B4b1a2
PHL110	Filipino	B4b1	B4b1	B4b1	B4b1a2
PHL142	Filipino	B4b1a2b	B4b1a2b	B4b1a2b	B4b1a2b1
PHL145	Filipino	B4	B4a1	B4a1	B4a1a5
PHL154	Filipino	B4b1	B4b1a+207	B4b1a+207	B4b1a2
SDNA035	S. Dakota Native American	B4	B4	B4b	B2c
SDNA060	S. Dakota Native American	B4	B4	B4b	B2c
SDNA126	**S. Dakota Native American**	**B4**	**B4c1c**	**B4b**	**B2**
SDNA129	S. Dakota Native American	B4	B4	B4b	B2c
SDNA130	S. Dakota Native American	B4	B4	B4b	B2
TXHis033	Texas Hispanic	B2+16278	B2+16278	B2+16278	B2q
TXHis167	Texas Hispanic	B4	B4	B4b	B2f
WANA007	**Washington Native American**	**B2c2a**	**B2c2a**	**B2c2a**	**B2**
WANA037	Washington Native American	B4	B4	B4b	B2
WANA050	**Washington Native American**	**B2c2a**	**B2c2a**	**B2c2a**	**B2**
WANA062	Washington Native American	B4	B4	B4b	B2
WANA065	Washington Native American	B2	B2a	B2a	B2a1
OHHis103	Washington Native American	B4	B4	B2l	-
SDNA029	S. Dakota Native American	B4	B4	B2	-

TXHis135	Texas Hispanic	B4	B4	B2o	-
ILH097	Illinois Hispanic	B4	B4	B2	-
JPN063	Japanese	B4	B4c1c1	-	-
NYAS062	Asian American	B4c1b+16335	B4c1b2a2	-	-
NYAS078	Asian American	B4c1b+16335	B4c1b2a2	-	-
PHL052	Filipino	B4c1b+16335	B4c1b2a2	-	-
PHL106	Filipino	B4c1b+16335	B4c1b2a2	-	-
PHL140	Filipino	B4c1b+16335	B4c1b2a2	-	-
CHN007	Chinese	B4a4	-	-	-
CHN129	Chinese	B4b1	-	-	-
CHN157	Chinese	B4a4	-	-	-
ILH074	Illinois Hispanic	B2t	-	-	-
ILH084	Illinois Hispanic	B2s	-	-	-
JPN138	Japanese	B4b1b	-	-	-
JPN274	Japanese	B4c1a1a	-	-	-
JPN275	Japanese	B4b1a1	-	-	-
OHHis035	Ohio Hispanic	B2g1	-	-	-
OHHis116	Ohio Hispanic	B2c2a	-	-	-
SibA009	Siberian: Altai Mtns	B4b1a3a	-	-	-
SibA096	Siberian: Altai Mtns	B4b1a3a	-	-	-
SibYDe54	Siberian Yakut	B4b1a3a	-	-	-
SibYDy05	Siberian Yakut	B4a4	-	-	-
SibYM002	Siberian Yakut	B4a4	-	-	-
SibYO025	Siberian Yakut	B4a4	-	-	-
TXHis117	Texas Hispanic	B2x	-	-	-
WANA093	Washington Native American	B2a	-	-	-

Table 4.7 displays each sample with the variants present in the full haplotype, the precise haplogroup using full mitogenome data, any SNPs found in that haplogroup that are not found in the sample (Missing SNPs), and any additional polymorphisms found in the sample that do not define the precise haplogroup (Private SNPs). High levels of missed SNPs illuminate any haplogrouping errors that may occur while high levels of private SNPs may indicate the sample belongs on a more derived haplogroup branch. Samples precisely haplogrouped as B2 have higher levels of private SNPs than those belonging to a subtype of haplogroup B2.

Table 4.7: Precise Haplogroups, Haplotypes, Missed SNPs, and Private SNPs for Each Sample. Length variations in the HV2 polycytosine stretch were omitted.

Sample	Population	Precise Haplo-group	Full Haplotype	Missed SNPs	Private SNPs
CHN094	Chinese	B4g1a	73G, 146C, 152C, 263G, 523-, 524-, 750G, 1282A, 1438G, 2706G, 3918A, 4769G, 4961C, 5108C, 7028T, 7789A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9968T, 11719A, 12540G, 14766T, 14905A, 15090C, 15326G, 16181C, 16182C, 16183C, 16189C, 16213A, 16217C, 16261T, 16292T, 16519C		146C, 152C, 1282A, 3918A, 4961C, 12540G, 15090C
ILH012	Illinois Hispanic	B2	73G, 263G, 499A, 750G, 827G, 1438G, 2706G, 3537G, 3547G, 4769G, 4820A, 4977C, 5477M, 6473T, 7028T, 7642A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8764A, 8860G, 9950C, 10172A, 11177T, 11719A, 13590A, 14766T, 15205T, 15313C, 15326G, 15535T, 15781T, 16183C, 16189C, 16213C, 16217C, 16519C		3537G, 7642A, 8764A, 10172A, 15205T, 15313C, 15781T, 16213C
ILH017	Illinois Hispanic	B2a1	73G, 263G, 499A, 524.1A, 524.2C, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 5975G, 6179A, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10160T, 10256C, 10895G, 11177T, 11719A, 11812G, 13590A, 14766T, 15068Y, 15326G, 15535T, 15762R, 16111T, 16183C, 16189C, 16217C, 16483A, 16519C		5975G, 6179A, 10160T, 10256C, 11812G
ILH070	Illinois Hispanic	B2b	73G, 228A, 263G, 499A, 573.1C, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 6755A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 13590A, 14100T, 14215C, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16311C, 16519C		228A, 573.1C, 14100T, 14215C, 16311C
ILH071	Illinois Hispanic	B2c2a	73G, 146C, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 3547G, 4755C, 4769G, 4820A, 4977C, 6473T, 7028T, 7241G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8452R, 8702T, 8860G, 9950C, 11177T, 11719A, 13590G, 14757C, 14766T, 15326G, 15535T, 16182C, 16183C, 16189C, 16217C, 16319R, 16519C, 16566A		

ILH087	Illinois Hispanic	B2+16278	73G, 146C, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4336C, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8843C, 8860G, 9950C, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 16145A, 16182C, 16183C, 16189C, 16217C, 16278T, 16519C	G6755A	146C, 4336C, 8843C, 16145A
ILH097	Illinois Hispanic	B2	73G, 114T, 146C, 152C, 263G, 499A, 709A, 750G, 827G, 1438G, 2706G, 3547G, 3834A, 4769G, 4820A, 4977C, 5581G, 6473T, 6872G, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8394T, 8860G, 9950C, 11177T, 11719A, 12745T, 13590A, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16519C		114T, 146C, 152C, 709A, 3834A, 5581G, 6872G, 8394T, 12745T
JPN080	Japanese	B4b1a1a	73G, 199C, 202G, 207A, 263G, 499A, 750G, 827G, 1438G, 2706G, 2831A, 4117C, 4769G, 4820A, 6023A, 6413C, 7028T, 7664A, 8206A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 13590A, 14766T, 15236G, 15326G, 15535T, 16069Y, 16136C, 16183C, 16189C, 16217C, 16242T, 16284G, 16519C		16242T
JPN260	Japanese	B4c1b1a	73G, 150T, 211G, 263G, 455.1T, 523-, 524-, 709A, 750G, 1119C, 1438G, 2706G, 3083C, 3221G, 3497T, 4742C, 4769G, 5826C, 7028T, 7909T, 8167A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 13111C, 13879C, 14162A, 14766T, 15326G, 15346A, 15391T, 16111T, 16140C, 16154C, 16182C, 16183C, 16189C, 16217C, 16274A, 16452C, 16497G, 16519C	15148A	3083C, 13111C, 16452C
OHHis068	Ohio Hispanic	B2k	73G, 146C, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4371C, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 13590A, 14311Y, 14766T, 15326G, 15535T, 16182C, 16183C, 16189C, 16217C, 16278T, 16519C		16278T
PHL012	Filipino	B4b1a2	73G, 204Y, 207A, 244G, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6216C, 6413C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16183C, 16189C, 16217C, 16519C		244G
PHL109	Filipino	B4b1a2	73G, 207A, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6216C, 6413C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9449T, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16182, 16183C, 16189C, 16217C, 16519C		9449T
PHL110	Filipino	B4b1a2	73G, 152C, 263G, 499A, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6216C, 6413C, 7028T, 7775A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16182C, 16183C, 16189C, 16217C, 16519C	207A	152C, 7775A
PHL142	Filipino	B4b1a2b1	73G, 207A, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6216C, 6413C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9305A, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16182C, 16183C, 16189C, 16217C, 16300G, 16519C		
PHL145	Filipino	B4a1a5	73G, 146C, 263G, 523-, 524-, 750G, 1438G, 2706G, 4048A, 4769G, 5465C, 6719C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 8865R, 9123A, 10172A, 10238C, 11719A, 12239T, 14766T, 15326G, 15481A, 15746G, 16182C, 16183C, 16189C, 16217C, 16261T, 16391A, 16519C		10172A, 15481A, 16391A

PHL154	Filipino	B4b1a2	73G, 207A, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6216C, 6413C, 6465A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9938Y, 9449T, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16182C, 16183C, 16189C, 16217C, 16519C		6465A, 9449T
SDNA035	S. Dakota Native American	B2c	73G, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 7241G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10646A, 11177T, 11719A, 11963A, 12810G, 13590A, 14766T, 15326G, 15535T, 16048A, 16104T, 16181G, 16182C, 16183C, 16189C, 16217C, 16519C	16181M	10646A, 11963A, 12810G, 16048A, 16104T, 16181G
SDNA060	S. Dakota Native American	B2c	73G, 263G, 499A, 709A, 750G, 827G, 1438G, 2414A, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 7241G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 11848T, 13590A, 14766T, 15326G, 15461Y, 15535T, 16182C, 16183C, 16189C, 16217C, 16519C		709A, 2414A, 11848T
SDNA126	S. Dakota Native American	B2	73G, 150T, 263G, 499A, 750G, 827G, 1438G, 2706G, 3321T, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 9967C, 11177T, 11719A, 13590A, 14766T, 15172A, 15326G, 15535T, 16183C, 16189C, 16217C, 16519C		150T, 3321T, 9967C, 15172A
SDNA129	S. Dakota Native American	B2c	73G, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 7241G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10646A, 11177T, 11719A, 11963A, 12810G, 13590A, 14766T, 15326G, 15535T, 16048A, 16104T, 16181G, 16182C, 16183C, 16189C, 16217C, 16519C		10646A, 11963A, 12810G, 16048A, 16104T, 16181G
SDNA130	S. Dakota Native American	B2	64T, 73G, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 3675G, 4769G, 4820A, 4977C, 5824A, 5899.1C, 6473T, 6719C, 7028T, 7948T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9770C, 9950C, 11177T, 11719A, 11890G, 13590A, 14766T, 15326G, 15535T, 16183C, 16184A, 16189C, 16217C, 16260T, 16519C		64T, 3675G, 5824A, 5899.1C, 6719C, 7948T, 9770C, 11890G, 16184A, 16260T
TXHis033	Texas Hispanic	B2q	73G, 146C, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 3866C, 4047C, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9287A, 9950C, 11041T, 11177T, 11719A, 12633T, 13590A, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16278T, 16519C		146C, 3866C, 9287A, 11041T, 12633T, 16278T
TXHis167	Texas Hispanic	B2f	73G, 263G, 499A, 524.1A, 524.2C, 750G, 827G, 1438G, 2706G, 3547G, 3796G, 3996T, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10535C, 11177T, 11719A, 13590A, 13833G, 13967T, 14766T, 14803T, 15326G, 15535T, 16183C, 16189C, 16217C, 16519C		13967T, 14803T
WANA007	Washington Native American	B2	73G, 146C, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4117C, 4734T, 4769G, 4820A, 4977C, 6473T, 7028T, 8020A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10084C, 11177T, 11719A, 13590A, 13801G, 14766T, 15326G, 15535T, 16148T, 16183C, 16189C, 16217C, 16319A, 16519C		146C, 4117C, 4734T, 8020A, 10084C, 13801G, 16148T, 16319A
WANA037	Washington Native American	B2	55.1T, 57C, 73G, 263G, 499A, 750G, 827G, 1282A, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 15940C, 16160G, 16183C, 16189C, 16217C, 16519C		55.1T, 57C, 1282A, 15940C, 16160G

WANA050	Washington Native American	B2	73G, 146C, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4117C, 4734T, 4769G, 4820A, 4977C, 6473T, 7028T, 7844G, 8020A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10084C, 11177T, 11719A, 13590A, 13801G, 14766T, 15326G, 15535T, 16148T, 16183C, 16189C, 16217C, 16319A, 16519C		146C, 4117C, 4734T, 7844G, 8020A, 10084C, 13801G, 16148T, 16319A
WANA062	Washington Native American	B2	73G, 263G, 499A, 750G, 827G, 986A, 1438G, 2706G, 3547G, 3585T, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 11809C, 13098G, 13590A, 14766T, 15326G, 15535T, 16153A, 16182C, 16183C, 16189C, 16217C, 16254R, 16390A, 16519C		986A, 3585T, 11809C, 13098G, 16153A, 16390A
WANA065	Washington Native American	B2a1	61-, 62-. 68A, 71.1G, 71.2G, 73G, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9612A, 9950C, 10895G, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 16111T, 16183C, 16189C, 16217C, 16249C, 16483A, 16519C		61-, 62-, 68A, -71.1G, -71.2G, 9612A, 16249C
OHHis103	Washington Native American	B2l	73G, 210G, 263G, 480C, 499A, 592T, 750G, 1438G, 1462A, 1979Y, 2706G, 3426G, 3547G, 4769G, 4820A, 4977C, 5147A, 6473T, 7028T, 7424G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 11914A, 13590A, 14766T, 15055Y, 15106R, 15301A, 15326G, 15535T, 16183C, 16189C, 16195-, 16217C, 16258C, 16263C, 16422C, 16438A, 16465T, 16519C	827G	210G, 480C, 592T, 1462A, 3426G, 5147A, 7424G, 11914A, 15301A, 16195-, 16258C, 16263C, 16438A, 16465T
SDNA029	S. Dakota Native American	B2	44-, 46C, 47A, 73G, 146C, 263G, 499A, 546G, 750G, 827G, 1438G, 2706G, 3290C, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 8973G, 9950C, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16266A, 16519C		44-, 46C, 47A, 146C, 546G, 3290C, 8973G, 16266A
TXHis135	Texas Hispanic	B2o	73G, 146C, 150T, 151T, 263G, 499A, 750G, 827G , 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10914A, 11177T, 12136C, 12142G, 13590A, 14766T, 15326G, 15535T, 16092C, 16104T, 16183C, 16189C, 16217C, 16519C	G11719A	146C, 150T, 151T, 10914A, 12136C, 12142G, 12693G, 16104T
ILH097	Illinois Hispanic	B2	73G, 114T, 146C, 152C, 263G, 499A, 709A, 750G, 827G, 1438G, 2706G, 3547G, 3834A, 4769G, 4820A, 4977C, 5581G, 6473T, 6872G, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8394T, 8860G, 9950C, 11177T, 11719A, 12745T, 13590A, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16519C	1438G	279C, 1944T, 9438A
JPN063	Japanese	B4c1c1	73G, 150T, 189R, 195C, 200G, 214G, 263G, 750G, 1119C, 1438G, 2706G, 3497T, 4769G, 5441G, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 10398G, 11719A, 13629G, 14178C, 14766T 15326G, 15346A, 15941C, 16048A, 16183C, 16189C, 16217C, 16304C, 16311C		200G, 14178C, 16048A, 16304C
NYAS062	Asian American	B4c1b2a2	73G, 146C, 150T, 195C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 3571T, 4769G, 6383A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8772C, 8860G, 11299C, 11719A, 13105G, 13708A, 14766T, 15301A, 15326G, 15346A, 16140C, 16182C, 16183C, 16189C, 16217C, 16274A, 16335G, 16400T, 16519C		6383A, 11299C, 13105G, 13708A, 16400T
NYAS078	Asian American	B4c1b2a2	73G, 146C, 150T, 195C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 3571T, 4769G, 6383A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8772C, 8860G, 8994R, 11719A, 13105G, 13708A, 14766T, 15301A, 15326G, 15346A, 16140C, 16182C, 16183C, 16189C, 16217C, 16274A, 16335G, 16519C		6383A, 13105G, 13708A

PHL052	Filipino	B4c1b2a2	73G, 146C, 150T, 195C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 3571T, 4769G, 6383A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8772C, 8860G, 11299C, 11719A, 13105G, 13708A, 14766T, 15301A, 15326G, 15346A, 16140C, 16182C, 16183C, 16189C, 16217C, 16274A, 16335G, 16519C		6383A, 11299C, 13105G, 13708A
PHL106	Filipino	B4c1b2a2	73G, 146C, 150T, 152C, 195C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 3571T, 4769G, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8772C, 8860G, 9962A 11719A, 14766T, 15301A, 15326G, 15346A, 15497A, 16140C, 16182C, 16183C, 16189C, 16217C, 16274A, 16335G, 16519C		152C, 9962A, 15497A
PHL140	Filipino	B4c1b2a2	73G, 146C, 150T, 153G, 195C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 3571T, 4769G, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8745G, 8772C, 8860G, 11719A, 14766T, 15301A, 15326G, 15346A, 16140C, 16182C, 16183C, 16189C, 16217C, 16274A, 16335G, 16519C		153G, 8745G
CHN007	Chinese	B4a4	73G, 189G, 193G, 263G, 523-, 524-, 709A, 750G, 1438G, 2056A, 2706G, 4769G, 5465C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285-, 8286-, 8287, 8288-, 8289-, 8860G, 9123A, 9932A, 11719A, 13858G, 14133G, 14751T, 14766T, 15326G, 16182C, 16183C, 16189C, 16217C, 16261T, 16299G, 16519C		189G, 709A, 2056A, 9932A, 13858G, 14133G
CHN129	Chinese	B4b1	73G, 150T, 152C, 263G, 499A, 750G, 827G, 1438G, 1819C, 2706G, 4659A, 4688Y, 4769G, 4820A, 5301C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 13590A, 14766T, 15326G, 15535T, 16136C, 16182C, 16183C, 16189C, 16217C, 16270T, 16298C, 16519C		150T, 152C, 1819C, 4659A, 5301C, 16270T, 16298C
CHN157	Chinese	B4a4	73G, 152C, 193G, 263G, 523-, 524-, 709A, 750G, 1438G, 2706G, 3209T, 4769G, 5465C, 7028T, 7853A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 8889C, 9123A, 11719A, 14751T, 14766T, 15326G, 16182C, 16183C, 16189C, 16213A, 16217C, 16261T, 16295T, 16299G, 16519C		152C, 709A, 3209T, 7853A, 8889C, 16213A, 16295T
ILH074	Illinois Hispanic	B2t	73G, 263G, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10792G, 11177T, 11380G, 11719A, 13590A, 14766T, 15244G, 15326G, 15535T, 15884A, 16183C, 16189C, 16217C, 16259T, 16357C, 16467T, 16519C		11380G
ILH084	Illinois Hispanic	B2s	73G, 151T, 152C, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 3547G, 4679A, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8567C, 8860G, 9950C, 11177T, 11719A, 12616C, 13590A, 13740C, 14766T, 15326G, 15535T, 16152C, 16182C, 16183C, 16189C, 16217C, 16325C, 16519C		151T, 152C, 4679A
JPN138	Japanese	B4b1b	73G, 106-, 107-, 108-, 109-, 110-, 111-, 152C, 263G, 499A, 750G, 827G, 1391C, 1438G, 2706G, 4769G, 4820A, 5585A, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9554A, 11398T, 11719A, 13590A, 14569A, 14766T, 15326G, 15535T, 16136C, 16182C, 16183C, 16189C, 16217C, 16298C, 16362C, 16519C	3981G, 13934T, 16218T	106-, 107-, 108-, 109-, 110-, 111-, 9554A, 11398T
JPN274	Japanese	B4c1a1a	73G, 146C, 263G, 709A, 750G, 1119C, 1438G, 2706G, 3497T, 4769G, 5899.1C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 8873R, 10310A, 11719A, 12505T, 14133G, 14766T, 15326G, 15346A, 16086C, 16182C, 16183C, 16189C, 16217C, 16311C, 16519C		146C, 5899.1C, 12505T

JPN275	Japanese	B4b1a1	73G, 199C, 202G, 207A, 263G, 499A, 750G, 827G, 1438G, 2706G, 2831A, 4117C, 4769G, 4820A, 6023A, 6413C, 7028T, 8206A, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 11719A, 12121C, 13590A, 14766T, 15236G, 15326G, 15535T, 16093C, 16136C, 16183C, 16189C, 16217C, 16284G, 16390A		12121C, 16093C, 16390A,
OHHis035	Ohio Hispanic	B2g1	73G, 114G, 146C, 263G, 499A, 709A, 750G, 827G, 1002T, 1438G, 2706G, 3547G, 3766C, 4769G, 4820A, 4977C, 6164T, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 16183C, 16189C, 16217C, 16298C, 16519C		146C, 709A
OHHis116	Ohio Hispanic	B2c2a	73G, 146C, 263G, 499A, 523-, 524-, 750G, 827G, 1438G, 2706G, 3547G, 4755C, 4769G, 4820A, 4977C, 6473T, 7028T, 7241G, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8702T, 8860G, 9950C, 11177T, 11719A, 13590A, 14757C, 14766T, 15326G, 15535T, 16182C, 16183C, 16189C, 16217C, 16319A, 16519C		
SibA009	Siberian	B4b1a3a	73G, 146C, 207A, 263G, 408A, 499A, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 4822Y, 6023A, 6413C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9055A, 9338T, 9615C, 11719A, 13590A, 14133G, 14766T, 15326G, 15535T, 16086C, 16136C, 16182C, 16183C, 16189C, 16217C, 16519C		
SibA096	Siberian	B4b1a3a	73G, 146C, 207A, 263G, 408A, 499A, 524.1A, 524.2C, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6413C, 6524C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9055A, 9338T, 9615C, 11719A, 13590A, 14133G, 14766T, 15326G, 15535T, 15813C, 16086C, 16136C, 16183C, 16189C, 16217C, 16519C		6524C, 15813C
SibYDe54	Siberian	B4b1a3a	73G, 146C, 207A, 263G, 408A, 499A, 524.1A, 524.2C, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6413C, 7028T, 7511C, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9055A, 9338T, 9615C, 11719A, 13590A, 14133G, 14766T, 15326G, 15535T, 16086C, 16136C, 16183C, 16189C, 16217C, 16519C		7511C
SibYDy05	Siberian	B4a4	73G, 146C, 207A, 263G, 408A, 499A, 524.1A, 524.2C, 750G, 827G, 1438G, 2706G, 4769G, 4820A, 6023A, 6413C, 7028T, 7511C, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9055A, 9338T, 9615C, 11719A, 13590A, 14133G, 14766T, 15326G, 15535T, 15944C, 16086C, 16136C, 16183C, 16189C, 16217C, 16294T, 16519C		152C, 709A, 2222G, 4841A, 15944C, 16294T
SibYM002	Siberian	B4a4	73G, 152C, 193G, 263G, 709A, 750G, 1438G, 2222G, 2706G, 4769G, 4841A, 5465C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9123A, 11719A, 14751T, 14766T, 15326G, 15944C, 16182C, 16183C, 16189C, 16217C, 16261T, 16294T, 16299G, 16519C		152C, 709A, 2222G, 4841A, 15944C, 16294T
SibYO025	Siberian	B4a4	73G, 152C, 193G, 263G, 709A, 750G, 1438G, 2222G, 2706G, 4769G, 4841A, 5465C, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-,8860G, 9123A, 11719A, 14751T, 14766T, 15326G, 15944C, 16182C, 16183C, 16189C, 16217C, 16261T, 16294T, 16299G, 16519C		152C, 709A, 2222G, 4841A, 15944C, 16294T
TXHis117	Texas Hispanic	B2x	73G, 263G, 318C, 499A, 750G, 827G, 1438G, 2706G, 3547G, 4129G, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8860G, 9950C, 10646A, 11177T, 11389T, 11719A, 12346T, 13590A, 14100T, 14766T, 15014Y, 15326G, 15535T, 16000A, 16183C, 16189C, 16217C, 16299G, 16323C, 16519C		318C, 14100T, 16000A, 16299G

WANA093	Washington Native American	B2a	73G, 151T, 263G, 499A, 750G, 827G, 954T, 1438G, 2706G, 3547G, 4310R, 4769G, 4820A, 4977C, 6473T, 7028T, 8281-, 8282-, 8283-, 8284-, 8285, 8286-, 8287-, 8288-, 8289-, 8843C, 8860G, 9950C, 11177T, 11719A, 13590A, 14766T, 15326G, 15535T, 16111T, 16183C, 16189C, 16217C, 16399G, 16483A, 16519C		151T, 954T, 8843C, 16399G

Discerning Native American from Asian Haplogroups

While the precise haplogroup was not always revealed without full mitogenome data, in some samples, sufficient resolution was provided to accurately distinguish Native American from Asian ancestry. As seen in Table 4.6, the full mitogenome was needed to distinguish Native American haplogroup B2 and B2 subgroups from Asian B4 subgroups in 18% (10/56) of the samples. Using HV1 alone, it was possible to distinguish between B2 and B4 haplogroups in 68% (38/56) of samples. Using HV1 and 2 data alone, an additional 5% (3/56) of samples could make the B2/B4 distinction. When the full CR data was used, an additional 9% (5/56) of samples could make the B2/B4 distinction.

Figure 4.1 displays the maximum parsimony phylogenetic tree constructed from B2 samples, identifying variants that define each haplogroup. All haplogroups produced from Empop (Parson and Dür, 2007) corresponded with the haplogroups identified by MtPhyl using Phylotree v17 (van Oven, 2015). All B2 samples contain the five haplogroup-defining polymorphisms: 3547G, 4977C, 6473T, 9950C, 11177T. A clear distinction is observed between B4b1 and B2, reinforcing the haplogrouping results identified by Empop (Parson and Dür, 2007). As shown, high levels of private polymorphisms and variation exist among samples belonging to B2 and B2 subtypes. Of the B2 samples, an average of 6.6 private polymorphisms per sample are observed while an average of 4.4 private polymorphisms per sample are observed in subtypes of B2.

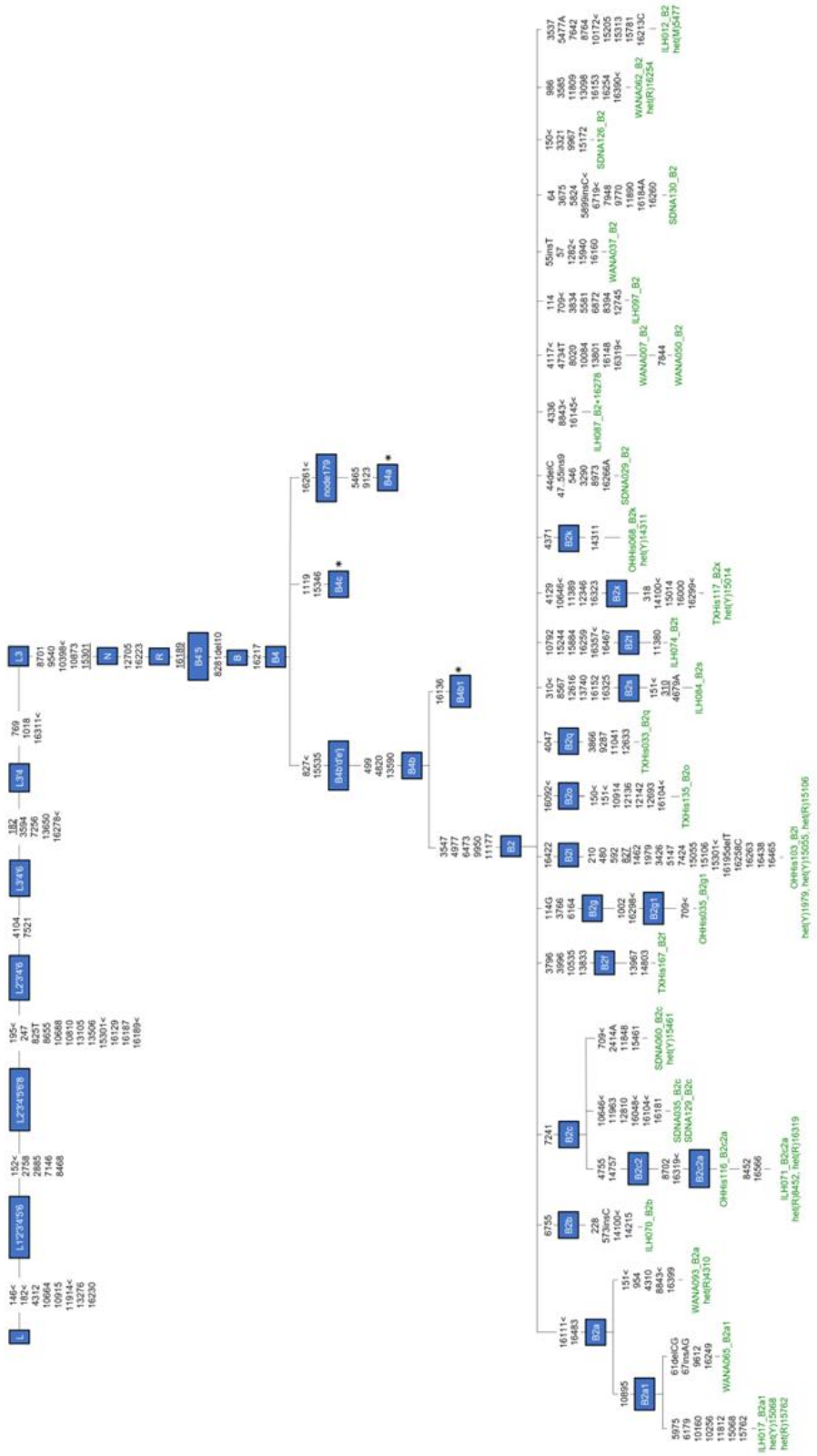


Figure 4.1: Phylogenetic tree of B2 and B2 Subtype Samples. Asterisks indicate phylogenetic trees represented in supplementary figures. Underlined variants indicate a back mutation to the ancestral state. Variants with '<' indicate positions that mutate numerous times within the mtDNA tree.

Discussion

Haplogrouping Accuracy

HV1 & 2, which are commonly used to characterize haplogroups, provide relatively high discriminating capabililties, however these cannot be used in all cases. Approximately one-sixth of the B4 haplogroup assignments examined here belong to B2 but could not be identified unless additional data were obtained from the full mitogenome. This is problematic if only the CR or HV regions are chosen for analysis and can result in an incorrect maternal ancestry prediction.

Caution should be exercised when estimating haplogroups from CR data alone as many of the defining variants are found in the mtDNA coding region. This is demonstrated in the data by three interesting cases, all Native American lineage assignments. The first (SDNA126) was assigned haplogroup B4c1c using HV1 & 2 data; however, this assignment reverted to B4b with a CR haplotype and B2 when the entire mitogenome data was considered. As B4c1c (Asian lineage) and B4b (both Asian and Native American lineage) are separate clades, this was an unexpected initial assignment. Using Phylotree v17 (van Oven, 2015), a closer examination revealed there are nine additional variants used to designate B4c1c from B4 but only two (150T and 195C) are located within the CR, specifically in HV2. This samples contains a 150T variant, resulting in the assignment of this haplogroup by EMPOP when only data from HV1 & 2 are used. In many population studies, haplogroup assessments are primarily determined based on HV1 and HV2 as they tend to be the most informative regions for human identification and more feasible with Sanger sequencing. If only these regions were sequenced, this sample would be incorrectly estimated as originating from Asia and not of indigenous American affiliation.

The second and third samples of note (WANA007 and WANA050) were assigned to haplogroup B2c2a using data from the CR but were ultimately identified as haplogroup B2 using

full mitogenome data. Closer examination revealed these sample had all variants necessary to designate them as B2 but only two of six additional variants needed to designate them as B2c2a. The two present variants (146C and 16319A) are opportunely located within HV1 and 2 while the four absent variants are located outside the CR, thus not detected without sequencing the full mitogenome. While these samples appropriately belong to B2, they were incorrectly assumed to be B2c2a when only HV1 and 2 data were included.

Examination of the phylogenetic tree in Figure 4.1 reveals high levels of private polymorphisms and variation among samples belonging to haplogroup B2 and B2 subtypes, illustrating a high level of genetic diversity within haplogroup B2. An average of 6.6 private polymorphisms per sample belonging to B2 were noted while an average of 4.4 private polymorphisms per sample are noted in subtypes of B2. This can be compared to an average of 3.8 private polymorphisms seen among complete mitogenomes belonging to subtypes of the Native American haplogroup C4c (Kashani et al., 2012). Several of the precisely haplogrouped B2 samples appear to represent divergent B2 haplotypes, expanding our knowledge of the known genetic variability seen within B2. The presence of numerous private polymorphisms beyond the B2 defining polymorphisms suggests new branches may exist within B2 that are not presently named.

Furthermore, 20% of the private polymorphisms seen in the precisely haplogrouped B2 samples are located within HV1. When both HV1 & 2 regions are examined, 30% of these private polymorphisms can be obtained. This reveals that not only can these private SNPs and the noted genetic variability be useful in phylogenetic reconstruction, but can also be useful for haplogrouping efforts of these haplotypes, particularly when examining the HVs and CR.

Discerning Native American from Asian Haplogroups

Nearly 75% of samples were geographically distinguishable using only HV1 & 2. This is promising as most anthropological studies utilize these two regions to make haplogroup assessments. However, the full mitogenome was needed in 18% to distinguish the Native American B2 haplogroup from its shared Asian/Native American B4b haplogroup. This can be problematic for anthropologists, particularly if no context is available for the sample in question.

Although many B2 haplotypes will need full mitogenome sequencing to confirm assignment, certain B2 subtypes may be determined without the extra measure of full mitogenome sequencing. As highlighted in an adaptation from Phylotree v17 in Fig. 5.1, many of the subtypes in haplogroup B2 have variants found within the CR, and even the HVSs. More than half of the B2 subtypes can be estimated within the B2 family from CR data alone. Few of these haplotypes have all the defining variants within the CR but many have approximately half of their variants within this region, giving sufficient information to assign a sample to that haplogroup.

B2	3547G	4977C	6473T	9950C	11177T					
B2a	16111T	16483A								
	B2a1	10895G								
		B2a1a	14766C							
			B2a1a1	12729G						
		B2a1b	3027C	12890T						
	B2a2	9097G								
	B2a3	551G	5054A							
	B2a4	16092C								
		B2a4a	228A							
			B2a4a1	3663G	106685A	16325C				
	B2a5	189G	5987T	11884G	13221G	16278T!				
B2b	6755A									
	B2b1	152C!	207A	1041G	1842G	4226C	4814T	16093C	16175G	
	B2b2	16145A								
		B2b2a	209C	3394C	6260A	9233C	10915C!	11968G	16320T	
	B2b3	13708A								
		B2b3a	152C!	271T	3918A	4232C	15784C	16249C	16312G	
	B2b4	(159C)	(195C)	8641G	9605T	11569C	15521A	(16189T!!)	(16239T)	(16353T)
B2c	7241G									
	B2c1	9098C								
		B2c1a	6722A							
		B2c1b	4435G	7262G	7822G					
		B2c1c	14063C							
	B2c2	146C!	4755C	14757C						
		B2c2a	8702T	16319A						
		B2c2b	152C!	9682C	13661G	16295T				
B2d	498d	4122G	4123G	8875C	9682C					

B2e	6119T	14049T				
B2f	3796G	3996T	10535C	13833G		
B2g	114*G*	3766C	6164T			
	B2g1	1002T	16298C			
	B2g2	7340A	11647T	11875C		
B2h	11821G	16468C				
B2i	6272G					
	B2i1	430C	485C	961C	16311C!	
	B2i2	470G	11611A	15077A		
		B2i2a	16207G			
			B2i2a1	10248C	16291T	
				B2i2a1a	4259T	12400G
				B2i2a1b	3843G	
		B2i2b	207A			
			B2i2b1	153G	16249C	
B2j	131C	183G	5270T	15924G	16166G	16361A
B2k	146C!	4371C				
B2l	16422C					
B2m	15766G	16164G	16519*A*			
B2n	4191G	6383A				
B2o	16092C					
	B2o1	7270C	16145A			
		B2o1a	152C!	14500G		
B2p	2380T	8222C	11696A			
B2q	4047C					
B2r	5899d					
B2s	310C	8567C	126616C	13740C	16152C	16325C
B2t	10792G	15244G	15884A	16259T	16357C	16467T
B2u	152C!	182T!!	3927G	5492C	8146G	16256T
B2v	7376T	15661T	16140C			
B2w	146C!	11950G	14569A	16270*G*	16278T!	
B2x	4129G	10646A	11389T	12346T	16323C	
B2y	16261T					
	B2y1	3480G				

Figure 4.2: Haplogroups and Defining Variants Found within Haplogroup B2. Adapted from Phylotree v17 (van Oven, 2015). Variants in gray are in the CR. Variants in parentheses indicate recurrent/unstable mutations or are yet uncertain based on current data. Variants with an exclamation mark indicate a back mutation while two exclamation marks indicate a double back mutation. Deleted variants are indicated with a 'd'. Italicized mutations indicate a transversion.

One approach to separate B2 sequences from B4 sequences without full mitogenome sequencing is to target distinguishing variants. As seen in Figure 4.1, the split between haplogroups B4 and B4b is designated by the following variants: 499A, 4820A, 13590A. Variant 499A is opportunely found in the CR, outside of HV1 & 2; however, the two other variants are found outside the CR. If these three variants are present in a sample, it belongs on the B4b branch and may be associated with the Americas. However, haplogroup B4b has two subclades: B4b1, found in Asia, and B2, found in the Americas. Utilizing CR data to look for the 499A variant is not sufficient to predict Native American ancestry for an individual as the haplotype may also belong to B4b1, an Asian-affiliated haplogroup. However, the detection of

an additional variant, 16136C, can permit this distinction. If both 499A and 16136C are present, it is likely a B4b1 haplotype, indicating Asian origin. When the 499A variant is detected but the 16136C is not present, the sample most likely belongs to a B2 haplogroup, indicating Native American origin. This provides a method for distinguishing between B2 and B4b1 using the CR alone.

Manual examination of the mitogenome using this method relies on the stability of the 16136 position within haplogroups B4b1 and B2. This remains an issue as there is an inadequate number of sequenced mitogenomes belonging to B4b1 and B2 to derive accurate rates of variability at this position. The closest estimation would be to use the average of fluctuation rates over all haplogroups, not just those with B4b1 and B2. Ongoing research is being conducted to calculate the fluctuation rates of positions throughout the CR. This is done by examining the non-dominant occurrences of variants at each position using a wide-range of haplogroups that can be confidently identified using CR data alone. Using 30,142 CR haplotypes, the 16136C variant was calculated to have a fluctuation rate of 0.058059% (W. Parson, personal communication). Of the haplogroups used for this calculation, 10 belong to subtypes of haplogroup B4 with 137 samples total. All 137 B4 samples had the 16136C. While this frequency may vary when only B4b1 and B2 samples are examined, this is the closest current estimate for the fluctuation of the 16136C variant.

Chapter 5: NGS using Short Tandem Repeats

Traditional Methods

Genetic methods of human identification rely primarily on genotyping of autosomal specific short tandem repeat (STR) loci. STRs, also known as microsatellites, were first recognized in the 1980s for their individualistic nature, likening them to a 'fingerprint' on a genomic level (Jeffreys et al., 1985). Because of this, short tandem repeats became standard in forensic identifications. STRs are 2-7 nucleotides in length which tandemly repeat from half a dozen to several dozen times (Butler, 2007). The number of repeats of each STR vary by individual. If numerous autosomal STR loci are examined, it is possible to positively match biological samples. The STR loci included for examination are highly polymorphic, unlinked sites found in non-coding regions of the genome (Budowle et al., 2001). Additionally, because of the nature of segregating chromosomes, it is possible to use autosomal STRS (aSTRs) for identifying parentage. As diploid organisms, humans have two copies of each chromosome: one they inherit from their mother and one from their father. Each aSTR locus will have two alleles and will match one of the alleles observed in each parent.

While STRs provide a means for human identification, they have also been studied in population genetics (Crawford and Beaty, 2013). Recent human history can be examined using aSTRs as they have high mutation rates, between 10^{-4} and 10^{-3} mutations per locus per generation (Rubicz et al., 2006). Previous work has identified particular loci with high diversity and variance that are useful in anthropological work (Budowle et al., 2005; Lim et al., 2007) though numerous autosomal STR loci are still used to characterize populations (Zlojutro et al., 2006; Dos Santos et al., 2009; He et al., 2017; Shrivastava et al., 2017).

NGS of STRs

For decades STR genotyping was performed using PCR amplification and capillary electrophoresis (CE) to determine the variable lengths of individual STRs. Using one or two PCR amplifications, the13 CODIS loci can be obtained quickly, cost-effectively, and with low DNA template input (Schanfield, 2007). However, PCR-CE does not provide nucleotide sequence information like NGS (Yang et al., 2014). Bornman et al. (2012) presented one of the initial systematic methods for performing high-throughput genotyping for forensic applications. Using short reads of 150 base pairs to sequence the 13 CODIS STRs and the amelogenin locus used for determining sex, they demonstrated that all loci could be accurately called from both individuals samples as well as mixed samples using NGS methodology. This instigated further exploration into the usage of NGS in STR markers.

As mentioned previously, not only can the length of STR alleles be identified but the exact nucleotide sequence can be revealed as well (Rockenbauer et al., 2014; Scheible et al., 2014). The sample profile examined in Figure 5.1 has 10 repeats for the first and second alleles at this locus. However, these alleles are not the same as there is an intra-repeat sequence variant, also known as an isoallele. One allele has a point mutation, changing the four base pair repeat from T-C-T-A to T-C-T-G on the ninth repeat. In this case, earlier electrophoretic methods would not detect this polymorphism.

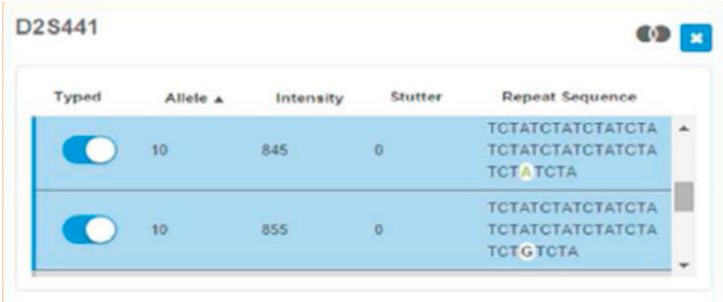


Figure 5.1: NGS sequencing of STRs Showing Variation in Sequences. STR is of same length but each repeat has a different variant. From Caratti et al., 2015.

Sequence data obtained through NGS methods also present new opportunities to identify variants that increase the statistical likelihood of a positive identification as an individual may have a unique polymorphism within the repeating STR allele (Børsting and Morling, 2015; Iozzi et al., 2015; Scheible et al., 2014). In a study of 3 aSTR loci (D3S1358, D12S391, D21S11) in 197 Danes, the match probability decreased from 0.0001 to 0.000005 when NGS was used over traditional PCR and capillary electrophoresis (Gelardi et al., 2014), demonstrating the additional discriminatory power NGS can offer in identification.

Y-chromosome STRs (Y-STRs) can be used when male samples are available. Y-chromosome STR length alleles are used in anthropological genetics to characterize populations (Mitchell et al., 2006; Crawford et al., 2010; Rubicz et al., 2010; Young et al., 2011; Wang et al., 2017) and track male migration (Ambrosio et al., 2010; Nuñez et al., 2010; Marks et al., 2012; Nagle et al., 2015; Olofsson et al., 2015). In forensic applications, they are often used in identification. One issue of using Y-STRs is that closely related males generally cannot be distinguished from one another since the male specific region of the Y chromosome is non-recombining and is passed down as a whole from father to son (Rubicz et al., 2006). However, NGS can help illuminate more variability by revealing sequence information not available using traditional methods.

Recent studies have revealed mutational rates in the Y chromosome at a rate of one mutation per generation (Xue et al., 2009; Xue and Tyler-Smith; 2010). Because of this, it has been suggested that theoretically, every Y chromosome is capable of being differentiated from one another. However, given that these tend to be single base substitutions (Xue et al., 2009), these mutations would not be seen using traditional PCR-CE methods. A frameshift mutation would have to occur to detect any changes as the length of the repeat would need to differ.

Using NGS, base substitutions from generational mutations can be learned and thus, related males can often be distinguished from one another. More recently, several rapidly mutating Y-STRs have been identified that can help distinguish between male relatives. These loci are capable of providing a 4.4-fold increase in discriminatory power of related males (Ballantyne et al., 2012). Because of this, these rapidly mutation loci are increasingly being used, particularly in identification settings.

Isoalleles in Population Genetics

To date, there have yet to be any anthropology-focused published works that examine isoalleleic rates across populations. The focus of STR isoalleles has been forensic applications of increased identification likelihoods (Gelardi et al., 2015; Scheible et al., 2014). However, several forensic investigations have suggested that isoalleles may be non-randomly distributed across populations and across all loci.

A study by van der Gaag et al. (2016) examined three distinct populations: Dutch, Himalayan, and Central African pygmies for 17 aSTR loci and noted significant variation that was not evenly dispersed over all loci. Sequence alleles as opposed to only length alleles provided a three-fold increase in discriminatory power for the D5S818 and D13S317 loci. Additionally, the D2S1338, D3S1358, D7S820, D8S1179, D16S539, and D21S11 loci all exhibited a two-fold difference in the match likelihood, revealing that some loci are more informative than others. When only the Himalayan and Dutch samples were examined, the D5S818, D7820, D13S317, and D13S317 loci all exhibit a greater than two-fold difference in match likelihood (van der Gaag et al., 2016), suggesting there is variation in the prevalence of sequence alleles between populations.

Novroski et al. (2016) also observed possible population structure among isoalleles of aSTRs when examining four major U.S. population groups. Departures from Hardy-Weinberg equilibrium (HWE) were detected at the D13S317, D5S818, and D7S820 loci in the U.S. Caucasian population group, at the D16S539 and D7S820 loci in the Hispanic population group, at the D13S317, and D16S539 loci in the African American population group, and at D7S820 in the Chinese population group. They conclude that these deviations might be the result of population substructure.

Planz et al. (2012) noted 11 of the 13 aSTR loci examined contained polymorphisms not observed by classical PCR-CE when examining three major U.S. population groups: African American, Caucasian, and Hispanic individuals. Seven of these loci showed a high degree of SNPs: D3S1358, D5S818, D7S820, D8S1179, D13S317, D21S11, and vWA. Further, an evaluation of population substructure was performed among the three population sets using length allele data versus length and sequence allele data using a two-sample Kolmogorov-Smirnov test. The locus specific values across the loci were comparable for all but two of the loci: vWA and D3S1358. It is determined there are underlying population-specific allele distributions at these two loci (Planz et al., 2012).

Gettings et al. (2016) examined 22 aSTR loci using a similar dataset to Planz et al. (2012). Six loci demonstrated more than twice the number of alleles obtained by sequence than by length alone: D1S391, D2S1338, D8S1179, vWA, and D3S1358. The last two loci, vWA, and D3S1358 were also pointed out by Planz et al. (2012) as loci that may have underlying population-specific distributions. Additionally, nine loci exhibited moderate gains in number of alleles obtained by sequence: D1S1656, D2S441, FGA, D18S51, Penta E, D19S433, D5S818,

CSF1PO, and D10S1248 while seven loci provided no additional alleles: Penta D, D22S1045, D13S317, D7S820, D16S539, TPOX, TH01.

Scheible et al. (2014) also noted a 21% increase in alleles by sequence, observing additional gains at the D12S391, D2S441, D3S1358, FGA, vWA, and D21S11 loci but found no gain for the D2S1338 locus. Zeng et al. (2015) found similar results for the D21S11 and D3S1358 loci but also added D2S1338, D8S1179, and vWA loci as well while Gelardi et al. (2014) observed an increase in alleles at loci D3S1358, D12S391, and D21S11. Zhao et al. (2016) noted four loci that demonstrated sequence allele variability: D3S1358, D2S441, D19S433, and D7S820. In addition, they observed variants in the flanking region outside the repeats at the D13S317, D16S539, D2S441, D5S818, D7S820, and TPOX loci. Finally, Novroksi et al. (2016) summarized all published datasets to view the most commonly reported loci with high sequence-based variance, noting three main loci: D2S1338, S12S391, and D21S11. These are consistent with previous findings discussed here. The overall findings of these studies are summarized below in Table 5.1

Table 5.1: aSTRs with Commonly Reported Sequence Allele Variation. "+" represents an increase in alleles with sequence data, "-" represents no increase in alleles with sequence data, "N/A" represents no data collected, "pop +" represents an increase in alleles with evidence of population distribution, and "flank +" represents an increase in alleles likely due to SNPs in the flanking regions.

Locus	Gettings et al., 2016	Planz et al., 2012	Zeng et al., 2015	Scheible et al., 2014	van der Gaag et al., 2016	Gelardi et al., 2014	Zhao et al., 2016	Novroski et al., 2016
D3S1358	+	pop +	+	+	+	+	+	+
D21S11	+	+	+	+	+	+	N/A	+
vWA	+	pop +	+	+	+	N/A	N/A	+
D5S818	+	+	-	flank +	pop +	N/A	flank +	pop +
D8S1179	+	+	+	-	+	N/A	N/A	+
D12S391	+	N/A	N/A	+	N/A	+	N/A	+

D2S1338	+	N/A	+	-	+	N/A	N/A	+
D7S820	-	+	-	flank +	pop +	N/A	+	pop +
FGA	+	-	-	+	+	N/A	N/A	+
D16S539	-	-	-	flank +	+	N/A	flank +	pop +
D2S441	+	N/A	N/A	+	N/A	N/A	+	+
D19S433	+	N/A	-	N/A	+	N/A	+	+
D13S317	-	+	-	-	pop +	N/A	flank +	pop +
D18S51	+	-	-	-	+	N/A	N/A	+
Penta E	+	N/A	-	N/A	+	N/A	N/A	+
TPOX	-	-	-	-	+	N/A	flank +	-
D10S1248	+	N/A	N/A	-	N/A	N/A	N/A	+
CSF1PO	+	-	-	-	-	N/A	N/A	+
D1S1656	+	N/A	N/A	-	N/A	N/A	N/A	+
Penta D	-	N/A	-	N/A	+	N/A	N/A	+
TH01	-	-	-	-	+	N/A	-	+
D9S1122	N/A	N/A	N/A	+	N/A	N/A	N/A	+
D16S539	N/A	N/A	N/A	+	N/A	N/A	N/A	+

Y-chromosome STRs have been examined by Warshauer et al. (2015) for 28 loci for the same three major U.S. groups as Planz and colleagues analyzed (2012). They reported 37 unique sequence alleles that had not been previously published as well as unique repeat pattern variants among complex STRs. Complex STRs are loci where the repeat motif consists of several repeating blocks with a different sequence (van der Gaag et a., 2016). Repeat pattern variants (RPVs) are described as allele sequences that differ from published data with regard to repeat unit arrangement (Warshauer et al., 2015). The structure remains consistent with the reported repeat motif but displays a pattern of repeat units that has not been published. This differs from SNP variants within repeats. For example, a length allele might contain 17 repeat units but there are several sequence variants and arrangements that can occur (Table 5.2).

Table 5.2: Example of Various STR Alleles that Share the Same Length.

Normal motif	**SNP within motif**	**Repeat Pattern Variant**
$[TCTA]_6[TCTG]_{11}$	$[TCTA]_5$**$[TATA]_1$**$[TCTG]_{11}$	$[TCTA]_5[TCTG]_{12}$

Interestingly, they noted that a particular RPV ($[TCTG]_6[TCTA]_{11}[TCTG]_3[TCTA]_9$) for the length allele "30" at locus DYS389II was observed in seven samples, all which belonged to African American individuals. Additionally, they observed a unique RPV ($[TCTG]_8[TCTA]_9[TCTG]_1[TCTA]_3$) for the length allele "21" at locus DYS390 in eight samples, all which belonged to African American individuals. This indicates these alternative allele sequences might be population-specific and further stresses a need for sample population characterization of STR alleles (Warshauer et al., 2015). This study is one of very few that takes into account not only sequence-based alleles but also repeat pattern variants.

Kwon et al. (2016) examined 23 Y-STR loci and observed sequence allele variation at the DYS19, DYS389I/II, DYS390, DYS392, DYS393, DYS437, DYS438, DYS448, DYS481, DYS635, and YGATAH4 loci, noting that DYS389II exhibited exceptional variation. Wendt et al. (2016) noted sequence alleles at the DYS389II and DYS448 loci, like Kwon et al. (2016), but also noted additional variation at the DYF387S1 and DYS390 loci. Zhao et al. (2015) also noted sequence variants at DYS389II and DYS448 like Wendt et al. (2016) and Kwon et al. (2016). Additionally, Zhao et al. (2015) observed sequence variation at DYS390 like Wendt et al. (2016) and at DYS437, DYS438, and DYS635 like Kwon et al. (2016).

Further, Novroksi et al (2016) noted four Y-STR loci that produced a 20% increase in total alleles when sequence-based alleles were included: DYF387S1, DYS448, DYS635, and DYS437 while the DYS385a-b locus produced less than a 20% increase. Additionally, several

loci had a 20% increase in total alleles when both sequence-based alleles and flanking region alleles were accounted for: DYS437, DYS481, DYS390, DYS522, DYS438, DYS385a-b, DYS533, and DYS389II. D'Amato et al. (2010) reported different alleles represented among European individuals as compared to two other distinct population groups at the DYS481 loci. The overall findings of these studies are summarized below in Table 5.3.

Table: 5.3: Y-STRs with Commonly Reported Sequence Allele Variation. "+" represents an increase in alleles with sequence data, "-" represents no increase in alleles with sequence data, "N/A" represents no data collected, "pop +" represents an increase in alleles with evidence of population distribution, and "flank +" represents an increase in alleles likely due to SNPs in the flanking regions.

Locus	Kwon et al., 2016	Wendt et al., 2016	Zhao et al., 2015	D'Amato et al., 2010	Novroski et al., 2016	Just et al., 2017
DYS389II	+	+	+	N/A	flank +	+
DYS448	+	+	+	N/A	+	-
DYS635	+	-	+	N/A	+	-
DYS437	+	-	+	N/A	+	-
DYS438	+	-	+	N/A	flank +	-
DYS390	-	+	+	N/A	flank +	-
DYF387S1	N/A	+	N/A	N/A	+	+
DYS392	+	-	-	N/A	+	-
DYS389I	+	-	-	N/A	-	-
DYS570	-	-	N/A	N/A	+	-
DYS576	-	-	N/A	N/A	+	-
DYS522	N/A	-	N/A	N/A	flank +	-
DYS19	+	-	-	N/A	-	-
Y-GATA-H4	+	-	-	N/A	-	-
DYS385a-b	-	-	N/A	N/A	+	-
DYS612	N/A	-	N/A	-	+	-
DYS460	N/A	-	N/A	N/A	+	-
DYS533	-	-	N/A	N/A	flank +	-
DYS549	-	-	N/A	N/A	-	-
DYS643	-	-	N/A	N/A	-	-
DYS505	N/A	-	N/A	N/A	-	-

DYS391	-	-	-	N/A	-	-
DYS439	-	-	-	N/A	-	-

Almost all of the studies examined here have advocated for population characterization of aSTR and Y-STR isoalleles. Sequence variants and repeat pattern variants are clearly of interest not only for forensic identification likelihoods, but also for anthropological inferences. This research will focus on aSTR and Y-STR loci with commonly reported sequence-based allele variants in an effort to identify if population structure exists among these loci that can be useful in anthropological genetics moving forward.

Chapter 6: Short Tandem Repeat Analyses

Materials & Methods

Samples

Of the larger collection described in Chapter 4, 74 samples were selected for examination of autosomal STRs (Table 6.1) and 35 samples were selected for examination of Y-STRs (Table 6.2). Appropriate controls were processed with all samples including NC, PC (2800 M, Promega, Madison, WI), and RBs.

Table 6.1: List of Samples for Autosomal STR Analyses.

Affiliation	No. of Samples
Chinese	2
Japanese	11
Filipino	21
Asian American	4
Siberian	5
Native American: South Dakota	13
Native American: Washington	4
Hispanic: South Dakota	2
Hispanic: Illinois	4
Hispanic: Ohio	4
Hispanic: Texas	4
Total:	74

Table 6.2: List of Samples for Y-STR Analyses.

Y-STR Samples	
Affiliation	**No. of Samples**
Japanese	7
Filipino	13
Asian American	2
Siberian	1
Native American: South Dakota	5
Hispanic: Illinois	1
Hispanic: South Dakota	2
Hispanic: Ohio	4
Total:	35

Amplification

A total of 76 samples were amplified (74 samples, one NC, and one PC) to obtain autosomal and Y-chromosome STRs using the ForenSeq™ DNA Signature Prep Kit (Illumina; San Diego, CA) and using Primer Mix B, according to the manufacturer's protocol (Illumina, 2015). The primer pairs in Primer Mix B target 27 autosomal STRs, Amelogenin, 7 X-STRs, 24 Y-STRs, 94 identity SNPs, 56 ancestry SNPs, and 22 phenotypic SNPs. A selection of these loci were chosen for analysis here.

Seven of the top autosomal STR loci identified with high sequence-based allelic diversity were selected: D2S441, D2S1338, D7S820, D12S391, D16S539, FGA, and vWA. Five of the top Y-STR loci identified with high sequence-based allelic diversity were selected: DYS390, DYS392, DYS438, DYS448, and DYS635. These specific loci were chosen because they have previously demonstrated higher isoallelic variance and possible population-based frequency distributions (D'Amato et al., 2010; Planz et al. 2012; Gelardi et al., 2014; Scheible et al. 2014; Zeng et al. 2015; Zhao et al., 2015; Gettings et al. 2016; Kwon et al., 2016; Novroski et al., 2016;

Wendt et al., 2016; van der Gaag et al., 2016; Zhao et al., 2016; Just et al., 2017). Additionally, these loci were successful in producing sufficient read coverage to accurately identify repeat pattern variants and isoallelic variation.

Samples were diluted to 1 ng/µL and added to 10 µL of master mix that included: 4.7 µL of the ForenSeq PCR1 Reaction Mix, 0.3 µL of the ForenSeq enzyme mix (FEM), and 5 µL of ForenSeq DNA Primer Mix B (DPMB). Thermal cycling conditions outlined in Table 6.3 were applied.

Table 6.3: List of Autosomal STRs and Y-STRs used.

List of targeted STRs	
Autosomal STRs	**Y-STRs**
D2S441	DYS390
D2S1338	DYS392
D7S820	DYS438
D12S391	DYS448
D16S539	DYS635
FGA	
vWA	

Table 6.4: Thermocycler Conditions for PCR1 of STR Amplification.

Thermocycler Conditions		
98º	3 minutes	
96º	45 seconds	8 cycles
80º	30 seconds	
54º	2 minutes	
68º	2 minutes	
96º	30 seconds	10 cycles
68º	3 minutes	
68º	10 minutes	
10º	∞	

Library Preparation & Sequencing

Following amplification of the target loci, unique combinations of provided index adapters (i5, i7; Illumina, 2015) were added and amplified through a second PCR reaction for bioinformatic separation of samples downstream. The following was added to each sample: 4 µL of index 1 (i7), 4 µL of index 2 (i5), and 27 µL of ForenSeq PCR2 Reaction Mix. Thermal cycling conditions outlined in Table 6.5 were applied.

Table 6.5: Thermocycler Conditions for PCR2 of STR Amplification.

Thermocycler Conditions		
98º	30 seconds	
98º	20 seconds	15 cycles
66º	30 seconds	
68º	90 seconds	
68º	10 minutes	
10º	∞	

Libraries were then purified 1X using the provided purification beads and were normalized according to the manufacturer's protocol (Illumina, 2015). Normalized libraries were pooled in equal volumes of 5 µL. Hybridization buffer and Human Sequencing Control (HSC) were added to 7 µL of the pooled libraries. Heat was applied for denaturation prior to sequencing, according to the manufacturer's protocol. The denatured library pool was loaded onto a MiSeq FGx v2 300 cycle reagent cartridge (Illumina; San Diego, CA). DNA cluster generation and sequencing by synthesis was performed on a MiSeq FGx Instrument

Data Analyses

All MiSeq FGx data were analyzed using the accompanying ForenSeq Universal Analysis Software (UAS; Illumina, 2016). On the Illumina MiSeq FGx, ~85% of STR genotype

errors occur from allele dropouts where heterozygotes are falsely typed as homozygotes, generally with low input samples (Sharma et al., 2017). To address this, allele reads were calculated for each locus of each sample to ensure all homozygotes had sufficiently high coverage to be accurately called a homozygote. Additionally, it is known that ~9% of genotype errors result from high stutter that are typed as a true allele (Sharma et al., 2017). To address this, the UAS sets a default minimum analytical threshold of 10 reads and a default interpretation threshold of 30 reads. Each locus of each sample was carefully reviewed for calls above 10X but below 30X to distinguish stutter from real alleles.

The intralocus balance threshold was set to the default 60%. Any genotypes with greater than a 60% imbalance were manually examined for appropriate read coverage before being called (>50X per allele) or left uncalled if the intralocus balance exceeded 75%. Further, the authenticity of isoalleles and repeat pattern variants were thoroughly examined before genotype calls were made to eliminate calls resulting from a sequencing error. Sequence alleles were re-oriented to reflect the new NGS STR nomenclature proposed by Parson et al. (2016) to match nomenclature used by the STR sequence allele database compiled by Novroski et al. (2016).

Statistical Analyses

The number of length-based (LB) alleles, sequence-based (SB) alleles, and number of reads per locus were calculated using the ForenSeq UAS program and Microsoft Excel. The difference in number of LB and SB alleles were calculated to reveal differences in the output data from NGS as opposed to traditional PCR-CE. LB and SB allele frequencies were calculated per locus for the two main geographic separations in the data: continental Asia (AS) and the Americas (NA). These frequencies were compared with frequencies calculated from compiled datasets found in Novroski et al. (2016) for four main groups used in U.S. forensic datasets:

African American (AFA), Asian (ASN), Caucasian (CAU), and Hispanic (HIS). Novroski and colleagues (2016) note that these data, to the best of their knowledge, consist of all published sequence-based STR alleles in the ForenSeq panel to date.

Observed and expected heterozygosities were calculated and Hardy-Weinberg equilibrium (HWE) tests performed for both LB and SB alleles for the autosomal STR loci of both the NA and AS groups using the Arlequin 3.5 (Excoffier and Lischer, 2010) to test for non-random associations of alleles. This is a common statistical measure used in population genetics. Hardy-Weinberg law states that in a large random-mating population with no evolutionary factors like selection, mutation, genetic drift, or migration, the genotypic frequencies will remain constant from generation to generation (Guo and Thompson, 1992). Further, Hardy-Weinberg tests can illuminate excess homozygosity resulting from allelic dropout that can occur when using panels with numerous primer sets. The genotypic array for an *m*-allele autosomal locus with alleles $A_1, A_2, \ldots, A_m$ as given by the Hardy-Weinberg law is:

$$\sum_i p_i^2 A_i A_i + \sum_{i<j} 2 p_i p_j A_i A_j,$$

whereby p_i represents the allelic frequency of A_i. A population with these genotypic frequencies is in Hardy-Weinberg equilibrium at the locus of interest (Guo and Thompson, 1992). The exact test for Hardy-Weinberg proportions can be performed by first determining the probability of the sample $\boldsymbol{f}$:

$$\Pr(\mathbf{f}) = \frac{n! \prod \frac{m}{i=1} f_i!}{(2n)! \prod_{j>i} f_{ij}!} 2^{\Sigma_{j>i} f_{ij}}$$

Then, the exact test for Hardy-Weinberg proportions given the observed sample **f** has to evaluate:

$$P = \sum_{\mathbf{g} \in \varphi} \Pr(\mathbf{g}),$$

where $\varphi = \{\mathbf{g}: \Pr(\mathbf{g}) \leq \Pr(\mathbf{f}), \mathbf{g} \in \Gamma_0\}$, and

$\Gamma_0 = \Gamma(\mathbf{f}) = \{\mathbf{g}: \mathbf{g}$ has the same allele counts as $\{g_i\}$ as does $\mathbf{f}\}$.

Rejection or acceptance of the null hypothesis is dependent upon whether *P* is smaller tha.;n the significance level α (Guo and Thompson, 1992). P-values were estimated using the Markov Chain method. A Markov Chain is a sequence of random variables $\{X_i : i = 0, 1, 2,\ldots\}$ with the Markov property that given the present state, the future and past states are independent for all measurable sets *A* in *X* (Liang et al., 2010)*:*

$$\Pr(X_{t+1} \in A | X_0 = x_0, \ldots, X_t = x_t) = \Pr(X_{t+1} \in A | X_t = x_t)$$

This method is a sampling method based on permutations of switches of alleles per cell with each permutation (Yuan and Bonney, 3003). Fisher's exact test was performed for all loci using Genepop 4.2. This was performed to compare the frequency distributions of LB and SB alleles for the NA and AS groups to test for statistical differences. This test is recommended when samples sizes are relatively small or some cell frequencies are small or zero, as is the case with the current data (Guo and Thompson, 1992). Fisher's exact test compares two-way tables assuming the row and column classifications are independent and is computed by:

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

where *a, b, c,* and *d* represent the observed cell frequencies (Sokal and Rohlf, 2012). Using this test, statistically significant differences in allele distributions between the AS and NA groups can be identified. This will test for any statistical differences in the frequency distributions of LB alleles versus SB alleles.

Results

Allele Frequencies

Table 6.6 displays the number of length-based (LB) alleles, the number of sequence-based (SB) alleles, and the difference between the number of these alleles for each aSTR locus. This was calculated for both NA and AS groups together as well as separately. The D12S391 locus displayed the highest level of sequence-based alleles, with an increase in 15 alleles. The D2S1338 and vWA locus have an increase of 15 and 12 alleles, respectively. The D2S441 and D7S820 loci have a small increase in the number of sequence-based alleles while the FGA and D16S539 loci display no increase in alleles.

Table 6.6: Number of Unique Alleles Observed for aSTR Loci.

Observed Sequence Variation			
Loci	# of LB Alleles	# of SB Alleles	Difference
D2S441	**9**	**12**	**+3**
NA	7	9	+2
AS	8	11	+3
D2S1338	**11**	**27**	**+15**
NA	11	16	+5
AS	10	20	+10
D12S391	**13**	**32**	**+19**
NA	13	20	+7
AS	10	27	+17
D7S820	**8**	**9**	**+1**
NA	8	8	0
AS	6	7	+1
FGA	**14**	**14**	**0**
NA	10	10	0
AS	12	12	0
vWA	**7**	**19**	**+12**
NA	6	9	+3
AS	7	16	+9
D16S539	**7**	**7**	**0**
NA	6	6	0
AS	6	6	0

Table 6.7 displays the number of length-based (LB) alleles, the number of sequence-based (SB) alleles, and the difference between the number of these alleles for each Y-STR locus. This was calculated for both NA and AS groups together as well as separately. The DYS390 and DYS438 loci both observed an increase of three additional alleles, the DYS448 locus had an increase of 2 alleles, and the DYS635 locus had an increase of 1 allele when sequence-based alleles were included. Only the DYS392 locus had no increase in alleles when sequence-based alleles were included. The DYS448 and DYS635 loci had low numbers of successful coverage likely due to bioinformatic dropout for NA samples, thus, the number of additional alleles is likely underrepresented for these loci.

Table 6.7: Number of Unique Alleles Observed for Y-STR Loci.

Observed Sequence Variation			
Locus	# of LB Alleles	# of SB Alleles	Difference
DYS390	**6**	**9**	**+3**
NA	5	5	0
AS	5	8	+3
DYS392	**5**	**5**	**0**
NA	4	4	0
AS	5	5	0
DYS438	**5**	**8**	**+3**
NA	3	5	+2
AS	4	6	+2
DYS448	**4**	**6**	**+2**
NA	1	1	0
AS	4	6	+2
DYS635	**7**	**8**	**+1**
NA	3	3	0
AS	7	8	+1

Table 6.8 lists each LB and SB allele present in the dataset along with read coverage for the D2S441 locus. Length-based alleles with different sequence motifs are bolded the first time

they appear in the table to better display sequence differences. Nine LB alleles and 12 SB alleles are present in the dataset. The "10", "11", and "13" LB alleles all demonstrate two SB alleles.

Table 6.8: Observed LB and SB Alleles at D2S441 Locus.

D2S441				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
WANA050	NA	8	[TCTAT]6[TCTG]1[TCTA]1	686
JPN050	AS	9.1	A[TCTA]9	726
JPN260	AS	9.1	A[TCTA]9	764
JPN260	AS	9.1	A[TCTA]9	476
VTAS001	AS	9.1	A[TCTA]9	283
CHN007	AS	10	[TCTA]8[TCTG]1[TCTA]1	283
ILH087	NA	10	[TCTA]8[TCTG]1[TCTA]1	466
ILH087	NA	10	[TCTA]8[TCTG]1[TCTA]1	466
ILH097	NA	10	[TCTA]10	50
ILH097	NA	10	[TCTA]8[TCTG]1[TCTA]1	50
JPN200	AS	10	[TCTA]8[TCTG]1[TCTA]1	54
JPN242	AS	10	[TCTA]8[TCTG]1[TCTA]1	62
NYAS062	AS	10	[TCTA]8[TCTG]1[TCTA]1	755
NYAS078	AS	10	[TCTA]8[TCTG]1[TCTA]1	690
OHHis035	NA	10	[TCTA]8[TCTG]1[TCTA]1	535
OHHis103	NA	10	[TCTA]10	421
OHHis103	NA	10	[TCTA]8[TCTG]1[TCTA]1	205
OHHis116	NA	10	[TCTA]8[TCTG]1[TCTA]1	201
PHL035	AS	10	[TCTA]8[TCTG]1[TCTA]1	1633
PHL106	AS	10	[TCTA]8[TCTG]1[TCTA]1	1577
PHL106	AS	10	[TCTA]10	1129
PHL145	AS	10	[TCTA]8[TCTG]1[TCTA]1	901
SDHis001	NA	10	[TCTA]8[TCTG]1[TCTA]1	1327
SDNA029	NA	10	[TCTA]8[TCTG]1[TCTA]1	897
SDNA035	NA	10	[TCTA]10	784
SDNA060	NA	10	[TCTA]10	744
SDNA088	NA	10	[TCTA]8[TCTG]1[TCTA]1	819
SDNA126	NA	10	[TCTA]10	802
SDNA130	NA	10	[TCTA]10	2398
SDNA130	NA	10	[TCTA]8[TCTG]1[TCTA]1	2398
SibA096	AS	10	[TCTA]10	891
SibYO025	AS	10	[TCTA]10	804
TXHis117	NA	10	[TCTA]10	1381
TXHis135	NA	10	[TCTA]10	1222
WANA062	NA	10	[TCTA]8[TCTG]1[TCTA]1	85
WANA062	NA	10	[TCTA]8[TCTG]1[TCTA]1	18
WANA093	NA	10	[TCTA]10	246

CHN031	AS	11	[TCTA]11	94
JPN080	AS	11	[TCTA]11	1553
JPN138	AS	11	[TCTA]11	1044
JPN199	AS	11	[TCTA]11	481
JPN207	AS	11	[TCTA]11	398
JPN275	AS	11	[TCTA]11	326
OHHis068	NA	11	[TCTA]11	382
PHL050	AS	11	[TCTA]11	756
PHL050	AS	11	[TCTA]11	581
PHL052	AS	11	[TCTA]11	4138
PHL055	AS	11	[TCTA]11	4138
PHL061	AS	11	[TCTA]11	669
PHL071	AS	11	[TCTA]11	672
PHL071	AS	11	[TCTA]11	754
PHL079	AS	11	[TCTA]11	604
PHL088	AS	11	[TCTA]11	3539
PHL097	AS	11	[TCTA]11	3539
PHL098	AS	11	[TCTA]9[TCTG]1[TCTA]1	1926
PHL100	AS	11	[TCTA]11	1877
PHL109	AS	11	[TCTA]11	605
PHL110	AS	11	[TCTA]11	556
PHL110	AS	11	[TCTA]11	873
PHL140	AS	11	[TCTA]11	723
PHL140	AS	11	[TCTA]11	2547
PHL154	AS	11	[TCTA]11	2547
PHL154	AS	11	[TCTA]11	871
SDHis001	NA	11	[TCTA]11	697
SDHis020	NA	11	[TCTA]11	1001
SDNA003	NA	11	[TCTA]11	1008
SDNA003	NA	11	[TCTA]11	1283
SDNA029	NA	11	[TCTA]11	1151
SDNA035	NA	11	[TCTA]11	1217
SDNA052	NA	11	[TCTA]11	1228
SDNA055	NA	11	[TCTA]11	1093
SDNA058	NA	11	[TCTA]11	1010
SDNA058	NA	11	[TCTA]11	595
SDNA060	NA	11	[TCTA]11	696
SDNA088	NA	11	[TCTA]11	432
SDNA126	NA	11	[TCTA]11	524
SDNA127	NA	11	[TCTA]11	432
SDNA127	NA	11	[TCTA]11	324
SDNA150	NA	11	[TCTA]9[TCTG]1[TCTA]1	1063
SDNA150	NA	11	[TCTA]11	1063
SibA009	AS	11	[TCTA]11	1824
SibYDe54	AS	11	[TCTA]11	1824
SibYDe54	AS	11	[TCTA]11	1270
SibYDy05	AS	11	[TCTA]11	1270
SibYO025	AS	11	[TCTA]9[TCTG]1[TCTA]1	1031
TXHis117	NA	11	[TCTA]9[TCTG]1[TCTA]1	872

TXHis135	NA	11	[TCTA]9[TCTG]1[TCTA]1	1869
TXHis167	NA	11	[TCTA]9[TCTG]1[TCTA]1	1869
TXHis167	NA	11	[TCTA]9[TCTG]1[TCTA]1	103
VTAS001	AS	11	[TCTA]9[TCTG]1[TCTA]1	110
VTAS016	AS	11	[TCTA]9[TCTG]1[TCTA]1	52
WANA037	NA	11	[TCTA]9[TCTG]1[TCTA]1	45
WANA037	NA	11	[TCTA]11	390
WANA050	NA	11	[TCTA]9[TCTG]1[TCTA]1	390
WANA093	NA	11	[TCTA]11	254
PHL088	AS	11.3	[TCTA]4[TCA]1[TCTA]7	223
JPN063	AS	11.3	[TCTA]4[TCA]1[TCTA]7	47
JPN274	AS	11.3	[TCTA]4[TCA]1[TCTA]7	83
NYAS062	AS	11.3	[TCTA]4[TCA]1[TCTA]7	296
PHL012	AS	11.3	[TCTA]4[TCA]1[TCTA]7	201
PHL084	AS	11.3	[TCTA]4[TCA]1[TCTA]7	156
PHL097	AS	11.3	[TCTA]4[TCA]1[TCTA]7	189
PHL098	AS	11.3	[TCTA]4[TCA]1[TCTA]7	754
PHL100	AS	11.3	[TCTA]4[TCA]1[TCTA]7	754
PHL142	AS	11.3	[TCTA]4[TCA]1[TCTA]7	190
PHL142	AS	11.3	[TCTA]4[TCA]1[TCTA]7	179
SDNA055	NA	11.3	[TCTA]4[TCA]1[TCTA]7	97
SibYDy05	AS	11.3	[TCTA]4[TCA]1[TCTA]7	60
CHN007	AS	12	[TCTA]12	129
JPN050	AS	12	[TCTA]12	128
JPN138	AS	12	[TCTA]12	181
JPN207	AS	12	[TCTA]12	157
JPN242	AS	12	[TCTA]12	276
JPN274	AS	12	[TCTA]12	276
JPN275	AS	12	[TCTA]12	532
OHHis116	NA	12	[TCTA]12	532
PHL012	AS	12	[TCTA]12	99
PHL052	AS	12	[TCTA]12	114
PHL061	AS	12	[TCTA]12	502
PHL084	AS	12	[TCTA]12	400
SDNA106	NA	12	[TCTA]12	218
SibA096	AS	12	[TCTA]12	182
TXHis033	NA	12	[TCTA]12	467
VTAS016	AS	12	[TCTA]12	467
OHHis035	NA	13	[TCTA]13	352
PHL145	AS	13	[TCTA]10[TTTA]1[TCTA]2	410
SibA009	AS	13	[TCTA]13	51
ILH012	NA	14	[TCTA]11[TTTA]1[TCTA]2	50
ILH012	NA	14	[TCTA]11[TTTA]1[TCTA]2	421
ILH071	NA	14	[TCTA]11[TTTA]1[TCTA]2	337
ILH071	NA	14	[TCTA]11[TTTA]1[TCTA]2	815

JPN063	AS	14	[TCTA]11[TTTA]1[TCTA]2	582
JPN080	AS	14	[TCTA]11[TTTA]1[TCTA]2	50
JPN199	AS	14	[TCTA]11[TTTA]1[TCTA]2	56
JPN200	AS	14	[TCTA]11[TTTA]1[TCTA]2	1149
NYAS078	AS	14	[TCTA]11[TTTA]1[TCTA]2	1149
OHHis068	NA	14	[TCTA]11[TTTA]1[TCTA]2	795
PHL005	AS	14	[TCTA]11[TTTA]1[TCTA]2	705
PHL005	AS	14	[TCTA]11[TTTA]1[TCTA]2	553
PHL035	AS	14	[TCTA]11[TTTA]1[TCTA]2	516
PHL055	AS	14	[TCTA]11[TTTA]1[TCTA]2	26
PHL079	AS	14	[TCTA]11[TTTA]1[TCTA]2	37
PHL109	AS	14	[TCTA]11[TTTA]1[TCTA]2	155
SDHis020	NA	14	[TCTA]11[TTTA]1[TCTA]2	25
SDNA052	NA	14	[TCTA]11[TTTA]1[TCTA]2	1526
SDNA106	NA	14	[TCTA]11[TTTA]1[TCTA]2	1526
TXHis033	NA	14	[TCTA]11[TTTA]1[TCTA]2	153
CHN031	AS	16	[TCTA]13[TTTA]1[TCTA]2	250

Table 6.9 details the frequency of each observed LB and SB allele for the D2S441 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.9: D2S441 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

D2S441									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
8	[TCTAT]6[TCTG]1[TCTA]1	0.00	1.00	1	0.00	0.00	0.00	0.00	0
9.1	A[TCTA]9	1.00	0.00	1	0.00	1.00	0.00	0.00	12
10	[TCTA]8[TCTG]1[TCTA]1	0.45	0.60	20	0.06	0.21	0.22	0.50	232
10	[TCTA]10	0.25	0.75	12	0.21	0.31	0.24	0.25	126
11	[TCTA]11	0.60	0.40	47	0.26	0.20	0.33	0.21	467
11	[TCTA]9[TCTG]1[TCTA]1	0.36	0.64	11	0.54	0.11	0.06	0.29	35
11.3	[TCTA]4[TCA]1[TCTA]7	0.92	0.08	13	0.16	0.31	0.37	0.17	95
12	[TCTA]12	0.81	0.19	16	0.44	0.35	0.07	0.14	135
13	[TCTA]13	0.50	0.50	2	0.25	0.50	0.00	0.25	4
13	[TCTA]10[TTTA]1[TCTA]2	1.00	0.00	1	0.50	0.03	0.33	0.13	30
14	[TCTA]11[TTTA]1[TCTA]2	0.55	0.45	20	0.29	0.15	0.32	0.24	351
16	[TCTA]13[TTTA]1[TCTA]2	1.00	0.00	1	0.33	0.00	0.33	0.33	3
Total		83	62	144	373	329	406	382	1490

Table 6.10 lists each LB and SB allele present in the dataset along with read coverage for the D2S1338 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Eleven LB alleles and 27 SB alleles are present in the dataset. The “16”, “23”, and “25” LB alleles demonstrate two SB alleles while the “18”, “19”, “20”, “21”, “22”, and “24” LB alleles demonstrate three SB alleles.

Table 6.10: Observed LB and SB Alleles at D2S1338 Locus.

D2S1338				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
JPN050	AS	16	[GGAA]9[GGCA]7	547
JPN080	AS	16	[GGAA]9[GGCA]7	102
SDNA150	NA	16	[GGAA]**10**[GGCA]**6**	81
ILH012	NA	17	[GGAA]11[GGCA]6	271
ILH012	NA	17	[GGAA]11[GGCA]6	271
ILH087	NA	17	[GGAA]11[GGCA]6	56
JPN063	AS	17	[GGAA]11[GGCA]6	319
JPN242	AS	17	[GGAA]11[GGCA]6	737
NYAS062	AS	17	[GGAA]11[GGCA]6	18
OHHis035	NA	17	[GGAA]11[GGCA]6	862
PHL012	AS	17	[GGAA]11[GGCA]6	517

PHL035	AS	17	[GGAA]11[GGCA]6	386
PHL061	AS	17	[GGAA]11[GGCA]6	440
PHL084	AS	17	[GGAA]11[GGCA]6	1028
PHL110	AS	17	[GGAA]11[GGCA]6	459
PHL145	AS	17	[GGAA]11[GGCA]6	437
SDHis001	NA	17	[GGAA]11[GGCA]6	180
SDHis001	NA	17	[GGAA]11[GGCA]6	180
SDHis020	NA	17	[GGAA]11[GGCA]6	93
SibA096	AS	17	[GGAA]11[GGCA]6	131
SibYDe54	AS	17	[GGAA]11[GGCA]6	258
JPN050	AS	18	[GGAA]12[GGCA]6	565
JPN063	AS	18	[GGAA]**11**[GGCA]**7**	227
OHHis116	NA	18	[GGAA]12[GGCA]6	237
PHL098	AS	18	[GGAA]11[GGCA]7	632
PHL106	AS	18	[GGAA]11[GGCA]7	243
PHL140	AS	18	[GGAA]12[GGCA]6	506
PHL142	AS	18	[GGAA]12[GGCA]6	308
PHL154	AS	18	[GGAA]11[GGCA]7	420
SDNA060	NA	18	[GGAA]12[GGCA]6	144
SibA009	AS	18	[GGAA]11[GGCA]7	532
VTAS016	AS	18	[GGAA]**13**[GGCA]**5**	395
CHN031	AS	19	[GGAA]12[GGCA]7	335
ILH097	NA	19	[GGAA]12[GGCA]7	38
JPN138	AS	19	[GGAA]12[GGCA]7	795
JPN200	AS	19	[GGAA]12[GGCA]7	431
JPN207	AS	19	[GGAA]12[GGCA]7	504
JPN274	AS	19	[GGAA]12[GGCA]7	793
NYAS062	AS	19	[GGAA]**13**[GGCA]**6**	28
NYAS078	AS	19	[GGAA]12[GGCA]7	125
OHHis103	NA	19	[GGAA]**11**[GGCA]**8**	134
PHL012	AS	19	[GGAA]12[GGCA]7	481
PHL052	AS	19	[GGAA]12[GGCA]7	715
PHL055	AS	19	[GGAA]12[GGCA]7	217
PHL079	AS	19	[GGAA]12[GGCA]7	554
PHL084	AS	19	[GGAA]12[GGCA]7	1089
PHL097	AS	19	[GGAA]12[GGCA]7	1533
PHL097	AS	19	[GGAA]12[GGCA]7	1533
PHL109	AS	19	[GGAA]12[GGCA]7	102
SDNA029	NA	19	[GGAA]12[GGCA]7	143
SDNA052	NA	19	[GGAA]12[GGCA]7	168
SDNA055	NA	19	[GGAA]12[GGCA]7	54

SDNA088	NA	19	[GGAA]12[GGCA]7	26
SDNA127	NA	19	[GGAA]12[GGCA]7	435
SDNA127	NA	19	[GGAA]12[GGCA]7	435
SibA009	AS	19	[GGAA]12[GGCA]7	362
SibYDe54	AS	19	[GGAA]13[GGCA]6	171
SibYO025	AS	19	[GGAA]12[GGCA]7	37
SibYO025	AS	19	[GGAA]12[GGCA]7	37
WANA062	NA	19	[GGAA]12[GGCA]7	307
JPN199	AS	20	[GGAA]13[GGCA]7	1448
JPN260	AS	20	[GGAA]13[GGCA]7	766
PHL035	AS	20	[GGAA]**14**[GGCA]**6**	309
PHL055	AS	20	[GGAA]13[GGCA]7	239
PHL088	AS	20	[GGAA]14[GGCA]6	470
PHL109	AS	20	[GGAA]13[GGCA]7	312
PHL110	AS	20	[GGAA]13[GGCA]7	419
PHL145	AS	20	[GGAA]13[GGCA]7	424
SDNA035	NA	20	[GGAA]10[GAAA]**1**[GGAA]2[GGCA]7	37
SDNA052	NA	20	[GGAA]13[GGCA]7	143
SDNA130	NA	20	[GGAA]13[GGCA]7	195
TXHis167	NA	20	[GGAA]10[GAAA]1[GGAA]2[GGCA]7	269
WANA037	NA	20	[GGAA]14[GGCA]6	28
WANA050	NA	20	[GGAA]10[GAAA]1[GGAA]2[GGCA]7	35
WANA093	NA	20	[GGAA]13[GGCA]7	53
ILH071	NA	21	[GGAA]14[GGCA]7	139
ILH087	NA	21	[GGAA]11[GAAA]**1**[GGAA]**2**[GGCA]**7**	55
ILH097	NA	21	[GGAA]14[GGCA]7	19
OHHis103	NA	21	[GGAA]2[GGAC]**1**[GGAA]**11**[GGCA]**7**	90
PHL071	AS	21	[GGAA]2[GGAC]**1**[GGAA]**12**[GGCA]**6**	464
WANA037	NA	21	[GGAA]14[GGCA]7	23
JPN138	AS	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	792
JPN199	AS	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	1373
PHL050	AS	22	[GGAA]2[GGAC]1[GGAA]**13**[GGCA]**6**	581
PHL088	AS	22	[GGAA]2[GGAC]1[GGAA]13[GGCA]6	434
PHL142	AS	22	[GGAA]2[GGAC]1[GGAA]**14**[GGCA]**5**	271
SDNA029	NA	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	80
SDNA058	NA	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	86
SDNA106	NA	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	72
SDNA130	NA	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	210
TXHis117	NA	22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	236

CHN007	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	383
ILH071	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	144
JPN080	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	119
JPN200	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	420
JPN242	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	484
JPN260	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	830
JPN274	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	821
JPN275	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	708
OHHis035	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	406
OHHis068	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	138
OHHis116	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	173
PHL005	AS	23	[GGAA]2[GGAC]1[GGAA]**14**[GGCA]**6**	497
PHL079	AS	23	[GGAA]2[GGAC]1[GGAA]14[GGCA]6	417
PHL100	AS	23	[GGAA]2[GGAC]1[GGAA]14[GGCA]6	296
PHL100	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	342
PHL140	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	384
PHL154	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	280
SDNA003	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	58
SDNA055	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	62
SDNA058	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	57
SDNA088	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	27
TXHis033	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	224
TXHis167	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	214
VTAS001	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	374
VTAS016	AS	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	222
WANA050	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	13
WANA062	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	111
WANA093	NA	23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	52
CHN007	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	308
CHN031	AS	24	[GGAA]2[GGAC]1[GGAA]**15**[GGCA]**6**	263
JPN207	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	414
JPN275	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	634
PHL005	AS	24	[GGAA]2[GGAC]1[GGAA]15[GGCA]6	406
PHL052	AS	24	[GGAA]2[GGAC]1[GGAA]15[GGCA]6	581
PHL061	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	324
PHL071	AS	24	[GGAA]2[GGAC]1[GGAA]**16**[GGCA]**5**	411
PHL106	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	108
SDHis020	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	42
SDNA003	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	56
SDNA126	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	210
SDNA126	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	210
SibA096	AS	24	[GGAA]2[GGAC]1[GGAA]16[GGCA]5	96

SibYDy05	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	347
TXHis033	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	221
TXHis117	NA	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	152
VTAS001	AS	24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	292
NYAS078	AS	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	30
PHL050	AS	25	[GGAA]2[GGAC]1[GGAA]**16**[GGCA]**6**	473
PHL098	AS	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	286
SDNA035	NA	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	16
SDNA060	NA	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	46
SDNA106	NA	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	65
SDNA150	NA	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	27
SibYDy05	AS	25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	220
OHHis068	NA	26	[GGAA]2[GGAC]1[GGAA]15[GGCA]8	57
TXHis135	NA	INC		
TXHis135	NA	INC		

Table 6.11 details the frequency of each observed LB and SB allele at the D2S1338 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.11: D2S1338 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

D2S1338									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
16	[GGAA]9[GGCA]7	2.20	0.00	2	0.00	0.80	0.20	0.00	5
16	[GGAA]10[GGCA]6	0.00	1.00	1	0.17	0.03	0.51	0.29	35
17	[GGAA]11[GGCA]6	0.61	0.39	18	0.16	0.10	0.44	0.31	206
18	[GGAA]11[GGCA]7	1.00	0.00	5	0.11	0.48	0.18	0.23	44
18	[GGAA]12[GGCA]6	0.60	0.40	5	0.14	0.23	0.44	0.20	71
18	[GGAA]13[GGCA]5	1.00	0.00	1	1.00	0.00	0.00	0.00	2
19	[GGAA]11[GGCA]8	0.00	1.00	1	0.50	0.33	0.00	0.17	6
19	[GGAA]12[GGCA]7	0.68	0.32	25	0.14	0.24	0.19	0.43	206
19	[GGAA]13[GGCA]6	1.00	0.00	6	0.66	0.07	0.12	0.15	41
20	[GGAA]13[GGCA]7	0.67	0.33	9	0.16	0.26	0.36	0.22	117
20	[GGAA]14[GGCA]6	0.67	0.33	3	0.40	0.10	0.40	0.10	10
20	[GGAA]10[GAAA]1[GGAA]2[GGCA]7	0.00	1.00	3	0.00	0.26	0.00	0.74	19
21	[GGAA]14[GGCA]7	0.00	1.00	3	0.39	0.19	0.26	0.16	31
21	[GGAA]2[GGAC]1[GGAA]12[GGCA]6	1.00	0.00	1	1.00	0.00	0.00	0.00	13
21	[GGAA]2[GGAC]1[GGAA]11[GGCA]7	0.00	1.00	1	0.76	0.00	0.10	0.14	21
22	[GGAA]2[GGAC]1[GGAA]12[GGCA]7	0.29	0.71	7	0.39	0.15	0.16	0.30	67
22	[GGAA]2[GGAC]1[GGAA]13[GGCA]6	1.00	0.00	2	0.71	0.21	0.00	0.07	14
22	[GGAA]2[GGAC]1[GGAA]14[GGCA]5	1.00	0.00	1	0.33	0.67	0.00	0.00	3
23	[GGAA]2[GGAC]1[GGAA]13[GGCA]7	0.48	0.52	25	0.13	0.33	0.22	0.33	187
23	[GGAA]2[GGAC]1[GGAA]14[GGCA]6	1.00	0.00	3	0.47	0.32	0.11	0.11	19
24	[GGAA]2[GGAC]1[GGAA]14[GGCA]7	0.54	0.46	13	0.21	0.34	0.25	0.19	154
24	[GGAA]2[GGAC]1[GGAA]15[GGCA]6	1.00	0.00	3	0.00	0.60	0.13	0.27	15
24	[GGAA]2[GGAC]1[GGAA]16[GGCA]5	1.00	0.00	2	0.00	0.00	0.00	0.00	0
25	[GGAA]2[GGAC]1[GGAA]15[GGCA]7	0.43	0.57	7	0.25	0.15	0.44	0.16	81
25	[GGAA]2[GGAC]1[GGAA]16[GGCA]6	1.00	0.00	1	0.22	0.67	0.11	0.00	9
26	[GGAA]2[GGAC]1[GGAA]15[GGCA]8	0.00	1.00	1	0.50	0.00	0.33	0.17	6
Total Samples		90	59	149	305	321	383	373	1382

Table 6.12 lists each LB and SB allele present in the dataset along with read coverage for the D7S820 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Eight LB alleles and nine SB alleles are present in the dataset. The "11" LB allele demonstrates two SB alleles seen only among one sample.

Table 6.12: Observed LB and SB Alleles at the D7S820 Locus.

D7S820				
Sample	Pop	Length	Sequence Motif	Reads
ILH012	NA	7	[TATC]7	109
CHN007	AS	8	[TATC]8	161
JPN050	AS	8	[TATC]8	254

JPN138	AS	8	[TATC]8	321
NYAS062	AS	8	[TATC]8	17
NYAS062	AS	8	[TATC]8	17
OHHis116	NA	8	[TATC]8	332
PHL035	AS	8	[TATC]8	257
PHL052	AS	8	[TATC]8	517
PHL055	AS	8	[TATC]8	255
PHL061	AS	8	[TATC]8	327
PHL106	AS	8	[TATC]8	449
PHL106	AS	8	[TATC]8	449
PHL154	AS	8	[TATC]8	355
SDHis020	NA	8	[TATC]8	32
SDNA029	NA	8	[TATC]8	115
SDNA150	NA	8	[TATC]8	45
SibA009	AS	8	[TATC]8	170
SibA096	AS	8	[TATC]8	48
VTAS001	AS	8	[TATC]8	330
VTAS016	AS	8	[TATC]8	179
WANA062	NA	8	[TATC]8	331
ILH097	NA	9	[TATC]9	29
JPN080	AS	9	[TATC]9	74
JPN275	AS	9	[TATC]9	283
PHL142	AS	9	[TATC]9	348
ILH012	NA	10	[TATC]10	35
ILH097	NA	10	[TATC]10	22
JPN063	AS	10	[TATC]10	128
JPN242	AS	10	[TATC]10	288
OHHis035	NA	10	[TATC]10	323
OHHis103	NA	10	[TATC]10	276
OHHis103	NA	10	[TATC]10	276
OHHis116	NA	10	[TATC]10	294
PHL005	AS	10	[TATC]10	438
PHL052	AS	10	[TATC]10	409
PHL100	AS	10	[TATC]10	277
PHL142	AS	10	[TATC]10	263
PHL145	AS	10	[TATC]10	293
SDHis020	NA	10	[TATC]10	26
SDNA035	NA	10	[TATC]10	29
SDNA058	NA	10	[TATC]10	192
SDNA058	NA	10	[TATC]10	192
SDNA088	NA	10	[TATC]10	64
SDNA088	NA	10	[TATC]10	64
SDNA106	NA	10	[TATC]10	112

SDNA106	NA	10	[TATC]10	112
SDNA127	NA	10	[TATC]10	49
SDNA127	NA	10	[TATC]10	49
SDNA130	NA	10	[TATC]10	288
SDNA130	NA	10	[TATC]10	288
SibYDe54	AS	10	[TATC]10	56
SibYDy05	AS	10	[TATC]10	110
SibYO025	AS	10	[TATC]10	15
TXHis033	NA	10	[TATC]10	138
TXHis135	NA	10	[TATC]10	20
TXHis167	NA	10	[TATC]10	439
TXHis167	NA	10	[TATC]10	439
VTAS016	AS	10	[TATC]10	214
WANA037	NA	10	[TATC]10	32
WANA050	NA	10	[TATC]10	32
WANA093	NA	10	[TATC]10	68
CHN007	AS	11	[TATC]11	143
CHN031	AS	11	[TATC]11	161
ILH071	NA	11	[TATC]11	172
ILH071	NA	11	[TATC]11	172
ILH087	NA	11	[TATC]11	27
ILH087	NA	11	[TATC]11	27
JPN050	AS	11	[TCTA]**9**[TCTG]**1**[TCTA]**1**	218
JPN063	AS	11	[TATC]11	145
JPN138	AS	11	[TATC]11	278
JPN199	AS	11	[TATC]11	211
JPN200	AS	11	[TATC]11	362
JPN207	AS	11	[TATC]11	258
JPN242	AS	11	[TATC]11	261
JPN260	AS	11	[TATC]11	332
NYAS078	AS	11	[TATC]11	125
NYAS078	AS	11	[TATC]11	125
OHHis035	NA	11	[TATC]11	300
OHHis068	NA	11	[TATC]11	231
PHL005	AS	11	[TATC]11	384
PHL012	AS	11	[TATC]11	474
PHL012	AS	11	[TATC]11	474
PHL035	AS	11	[TATC]11	214
PHL050	AS	11	[TATC]11	300
PHL055	AS	11	[TATC]11	182
PHL061	AS	11	[TATC]11	266
PHL071	AS	11	[TATC]11	399
PHL071	AS	11	[TATC]11	399

PHL079	AS	11	[TATC]11	251
PHL088	AS	11	[TATC]11	485
PHL097	AS	11	[TATC]11	820
PHL097	AS	11	[TATC]11	820
PHL098	AS	11	[TATC]11	448
PHL109	AS	11	[TATC]11	393
PHL109	AS	11	[TATC]11	393
PHL110	AS	11	[TATC]11	586
PHL110	AS	11	[TATC]11	586
PHL140	AS	11	[TATC]11	715
PHL140	AS	11	[TATC]11	715
SDHis001	NA	11	[TATC]11	60
SDNA003	NA	11	[TATC]11	172
SDNA003	NA	11	[TATC]11	172
SDNA029	NA	11	[TATC]11	129
SDNA035	NA	11	[TATC]11	33
SDNA052	NA	11	[TATC]11	259
SDNA052	NA	11	[TATC]11	259
SDNA055	NA	11	[TATC]11	125
SDNA055	NA	11	[TATC]11	125
SDNA060	NA	11	[TATC]11	76
SDNA126	NA	11	[TATC]11	62
SDNA150	NA	11	[TATC]11	26
SibA096	AS	11	[TATC]11	35
SibYDe54	AS	11	[TATC]11	49
SibYO025	AS	11	[TATC]11	12
TXHis033	NA	11	[TATC]11	104
TXHis117	NA	11	[TATC]11	266
TXHis135	NA	11	[TATC]11	22
WANA037	NA	11	[TATC]11	20
JPN200	AS	12	[TATC]12	360
JPN207	AS	12	[TATC]12	243
JPN260	AS	12	[TATC]12	264
JPN274	AS	12	[TATC]12	487
JPN274	AS	12	[TATC]12	487
PHL050	AS	12	[TATC]12	266
PHL079	AS	12	[TATC]12	214
PHL084	AS	12	[TATC]12	568
PHL084	AS	12	[TATC]12	568
PHL088	AS	12	[TATC]12	481
PHL098	AS	12	[TATC]12	393
PHL100	AS	12	[TATC]12	228
PHL154	AS	12	[TATC]12	300

SDHis001	NA	12	[TATC]12	31
SDNA126	NA	12	[TATC]12	84
SibA009	AS	12	[TATC]12	120
TXHis117	NA	12	[TATC]12	185
VTAS001	AS	12	[TATC]12	266
WANA050	NA	12	[TATC]12	12
WANA062	NA	12	[TATC]12	209
WANA093	NA	12	[TATC]12	45
CHN031	AS	13	[TATC]13	100
JPN080	AS	13	[TATC]13	77
JPN199	AS	13	[TATC]13	170
JPN275	AS	13	[TATC]13	256
OHHis068	NA	13	[TATC]13	188
PHL145	AS	13	[TATC]13	266
SibYDy05	AS	13	[TATC]13	143
SDNA060	NA	14	[TATC]14	26

Table 6.13 details the frequency of each observed LB and SB allele at the D7S820 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.13: D7S820 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

D7S820									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
7	[TATC]7	0.00	1.00	1	0.44	0.00	0.56	0.00	9
8	[TATC]8	0.76	0.24	21	0.30	0.20	0.34	0.16	220
9	[TATC]9	0.75	0.25	4	0.28	0.13	0.41	0.18	170
10	[TATC]10	0.31	0.69	36	0.33	0.15	0.24	0.28	408
11	[TATC]11	0.59	0.41	54	0.20	0.30	0.21	0.30	430
11	[TCTA]9[TCTG]1[TCTA]1	1.00	0.00	1	0.00	0.00	0.00	0.00	0
12	[TATC]12	0.71	0.29	21	0.20	0.26	0.26	0.28	249
13	[TATC]13	0.86	0.14	7	0.17	0.20	0.30	0.33	54
14	[TATC]14	0.00	1.00	1	0.17	0.50	0.33	0.00	6
Total		84	62	146	398	336	418	395	1547

Table 6.14 lists each LB and SB allele present in the dataset along with read coverage for the D12S391 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Thirteen LB alleles and 32 SB alleles are present in the dataset. The “17” and “19.3” LB alleles demonstrate two SB alleles. The “18”, “19”, and “23” LB alleles demonstrate three SB alleles. The “21”, “22”, and “24” LB alleles demonstrate four SB alleles. The “20” LB allele demonstrates five SB alleles, higher than any other LB allele of other loci.

Table 6.14: Observed LB and SB Alleles at D12S391 Locus.

D12S391			
Sample	**Length**	**Sequence Motif**	**Reads**
SDNA058	16	[AGAT]9[AGAC]6[AGAT]1	128
WANA093	16	[AGAT]9[AGAC]6[AGAT]1	51
CHN007	17	[AGAT]10[AGAC]6[AGAT]1	212
JPN242	17	[AGAT]**9**[AGAC]**7**[AGAT]1	688
NYAS078	17	[AGAT]10[AGAC]6[AGAT]1	32
PHL035	17	[AGAT]10[AGAC]6[AGAT]1	434
PHL109	17	[AGAT]10[AGAC]6[AGAT]1	365
PHL110	17	[AGAT]10[AGAC]6[AGAT]1	413
SDHis001	17	[AGAT]10[AGAC]6[AGAT]1	88

SDNA088	17	[AGAT]10[AGAC]6[AGAT]1	29
CHN031	18	[AGAT]11[AGAC]6[AGAT]1	214
ILH012	18	[AGAT]11[AGAC]6[AGAT]1	74
JPN080	18	[AGAT]11[AGAC]6[AGAT]1	98
JPN138	18	[AGAT]11[AGAC]6[AGAT]1	521
JPN200	18	[AGAT]11[AGAC]6[AGAT]1	488
JPN260	18	[AGAT]11[AGAC]6[AGAT]1	914
JPN274	18	[AGAT]11[AGAC]6[AGAT]1	1110
JPN274	18	[AGAT]11[AGAC]6[AGAT]1	1110
JPN275	18	[AGAT]11[AGAC]6[AGAT]1	502
OHHis103	18	[AGAT]11[AGAC]6[AGAT]1	136
PHL005	18	[AGAT]**10**[AGAC]**7**[AGAT]1	870
PHL012	18	[AGAT]11[AGAC]6[AGAT]1	462
PHL050	18	[AGAT]11[AGAC]6[AGAT]1	661
PHL052	18	[AGAT]11[AGAC]6[AGAT]1	693
PHL052	18	[AGAT]**12**[AGAC]**5**[AGAT]1	759
PHL088	18	[AGAT]11[AGAC]6[AGAT]1	647
PHL109	18	[AGAT]11[AGAC]6[AGAT]1	279
PHL140	18	[AGAT]11[AGAC]6[AGAT]1	1053
PHL140	18	[AGAT]11[AGAC]6[AGAT]1	1053
PHL154	18	[AGAT]11[AGAC]6[AGAT]1	500
SDHis001	18	[AGAT]11[AGAC]6[AGAT]1	56
SDHis020	18	[AGAT]11[AGAC]6[AGAT]1	84
SDNA029	18	[AGAT]11[AGAC]6[AGAT]1	111
SDNA035	18	[AGAT]11[AGAC]6[AGAT]1	52
SDNA055	18	[AGAT]11[AGAC]6[AGAT]1	99
SDNA058	18	[AGAT]10[AGAC]7[AGAT]1	143
SDNA060	18	[AGAT]11[AGAC]6[AGAT]1	62
SDNA126	18	[AGAT]11[AGAC]6[AGAT]1	154
TXHis117	18	[AGAT]11[AGAC]6[AGAT]1	175
TXHis167	18	[AGAT]11[AGAC]6[AGAT]1	178
VTAS001	18	[AGAT]11[AGAC]6[AGAT]1	425
SDNA150	18.3	[AGAT]1[GAT]1[AGAT]9[AGAC]7[AGAT]1	56
SibA009	18.3	[AGAT]1[GAT]1[AGAT]9[AGAC]7[AGAT]1	665
CHN007	19	[AGAT]**11**[AGAC]**8**	326
ILH012	19	[AGAT]12[AGAC]6[AGAT]1	72
ILH071	19	[AGAT]12[AGAC]6[AGAT]1	59
JPN063	19	[AGAT]12[AGAC]6[AGAT]1	389
JPN063	19	[AGAT]12[AGAC]6[AGAT]1	389
JPN207	19	[AGAT]**11**[AGAC]**7**[AGAT]1	457
JPN207	19	[AGAT]12[AGAC]6[AGAT]1	501

JPN260	19	[AGAT]12[AGAC]6[AGAT]1	627
OHHis035	19	[AGAT]11[AGAC]7[AGAT]1	235
OHHis068	19	[AGAT]12[AGAC]6[AGAT]1	164
PHL005	19	[AGAT]12[AGAC]6[AGAT]1	751
PHL050	19	[AGAT]12[AGAC]6[AGAT]1	442
PHL061	19	[AGAT]12[AGAC]6[AGAT]1	537
PHL100	19	[AGAT]11[AGAC]7[AGAT]1	386
PHL106	19	[AGAT]12[AGAC]6[AGAT]1	217
SDNA035	19	[AGAT]12[AGAC]6[AGAT]1	37
SDNA106	19	[AGAT]11[AGAC]7[AGAT]1	133
SDNA126	19	[AGAT]11[AGAC]7[AGAT]1	127
SDNA127	19	[AGAT]11[AGAC]7[AGAT]1	206
SDNA130	19	[AGAT]12[AGAC]6[AGAT]1	603
SDNA130	19	[AGAT]12[AGAC]6[AGAT]1	603
SibA096	19	[AGAT]11[AGAC]7[AGAT]1	64
SibA096	19	[AGAT]12[AGAC]6[AGAT]1	122
SibYDy05	19	[AGAT]12[AGAC]6[AGAT]1	197
TXHis033	19	[AGAT]11[AGAC]7[AGAT]1	58
TXHis117	19	[AGAT]11[AGAC]7[AGAT]1	227
TXHis167	19	[AGAT]12[AGAC]6[AGAT]1	110
VTAS001	19	[AGAT]12[AGAC]6[AGAT]1	347
WANA062	19	[AGAT]11[AGAC]7[AGAT]1	35
SDNA052	19.2	[AGAT]4[AT]1[AGAT]8[AGAC]6[AGAT]1	118
SDNA055	19.2	[AGAT]4[AT]1[AGAT]8[AGAC]6[AGAT]1	115
OHHis116	19.3	[AGAT]**5**[GAT]**1**[AGAT]**7**[AGAC]**6**[AGAT]1	309
WANA093	19.3	[AGAT]**1**[GAT]**1**[AGAT]**10**[AGAC]**7**[AGAT]1	36
JPN138	20	[AGAT]12[AGAC]7[AGAT]1	460
JPN199	20	[AGAT]12[AGAC]7[AGAT]1	840
JPN199	20	[AGAT]12[AGAC]7[AGAT]1	840
JPN242	20	[AGAT]**11**[AGAC]**8**[AGAT]1	585
NYAS078	20	[AGAT]12[AGAC]7[AGAT]1	23
OHHis068	20	[AGAT]**13**[AGAC]**6**[AGAT]1	179
OHHis103	20	[AGAT]12[AGAC]7[AGAT]1	90
PHL061	20	[AGAT]12[AGAC]7[AGAT]1	519
PHL071	20	[AGAT]12[AGAC]7[AGAT]1	422
PHL079	20	[AGAT]12[AGAC]7[AGAT]1	317
PHL097	20	[AGAT]12[AGAC]7[AGAT]1	683
PHL097	20	[AGAT]11[AGAC]8[AGAT]1	719
PHL098	20	[AGAT]13[AGAC]6[AGAT]1	421
PHL106	20	[AGAT]**14**[AGAC]**5**[AGAT]1	173
PHL145	20	[AGAT]12[AGAC]7[AGAT]1	448

PHL145	20	[AGAT]**12**[AGAC]**8**	473
SDNA052	20	[AGAT]12[AGAC]7[AGAT]1	124
SDNA060	20	[AGAT]13[AGAC]6[AGAT]1	57
SDNA088	20	[AGAT]12[AGAC]7[AGAT]1	26
SDNA127	20	[AGAT]13[AGAC]6[AGAT]1	121
SDNA150	20	[AGAT]13[AGAC]6[AGAT]1	79
SibYDy05	20	[AGAT]13[AGAC]6[AGAT]1	94
TXHis033	20	[AGAT]12[AGAC]7[AGAT]1	35
VTAS016	20	[AGAT]13[AGAC]6[AGAT]1	226
VTAS016	20	[AGAT]12[AGAC]7[AGAT]1	244
ILH071	21	[AGAT]**14**[AGAC]**6**[AGAT]1	56
JPN050	21	[AGAT]12[AGAC]8[AGAT]1	312
JPN275	21	[AGAT]12[AGAC]8[AGAT]1	410
OHHis035	21	[AGAT]**13**[AGAC]**8**	175
PHL012	21	[AGAT]**13**[AGAC]**7**[AGAT]1	403
PHL055	21	[AGAT]14[AGAC]6[AGAT]1	255
PHL088	21	[AGAT]12[AGAC]8[AGAT]1	502
PHL098	21	[AGAT]12[AGAC]8[AGAT]1	448
PHL142	21	[AGAT]12[AGAC]8[AGAT]1	263
SibYDe54	21	[AGAT]13[AGAC]7[AGAT]1	30
WANA062	21	[AGAT]13[AGAC]7[AGAT]1	16
OHHis116	22	[AGAT]**13**[AGAC]**9**	131
PHL035	22	[AGAT]13[AGAC]8[AGAT]1	393
PHL071	22	[AGAT]13[AGAC]8[AGAT]1	448
PHL084	22	[AGAT]**14**[AGAC]**7**[AGAT]1	234
PHL084	22	[AGAT]13[AGAC]8[AGAT]1	281
PHL100	22	[AGAT]13[AGAC]8[AGAT]1	295
PHL142	22	[AGAT]**12**[AGAC]**10**	245
PHL154	22	[AGAT]13[AGAC]9	338
SDNA106	22	[AGAT]13[AGAC]9	125
SibA009	22	[AGAT]13[AGAC]9	495
SibYO025	22	[AGAT]13[AGAC]9	33
SibYO025	22	[AGAT]13[AGAC]9	33
JPN200	23	[AGAT]**12**[AGAC]**11**	355
PHL055	23	[AGAT]**13**[AGAC]**10**	219
SDHis020	23	[AGAT]**13**[AGAC]**9**[AGAT]1	71
SDNA003	23	[AGAT]14[AGAC]9	109
SDNA029	23	[AGAT]14[AGAC]9	97
SibYDe54	23	[AGAT]14[AGAC]9	12
CHN031	24	[AGAT]**15**[AGAC]8[AGAT]1	178

JPN080	24	[AGAT]15[AGAC]8[AGAT]1	66
PHL110	24	[AGAT]14[AGAC]9[AGAT]1	273
SDNA003	24	[AGAT]**13**[AGAC]**11**	71
JPN050	25	[AGAT]16[AGAC]8[AGAT]1	260
PHL079	25	[AGAT]16[AGAC]8[AGAT]1	310
ILH087	INC	INC	
ILH087	INC	INC	
ILH097	INC	INC	
ILH097	INC	INC	
NYAS062	INC	INC	
NYAS062	INC	INC	
TXHis135	INC	INC	
TXHis135	INC	INC	
WANA037	INC	INC	
WANA037	INC	INC	
WANA050	INC	INC	
WANA050	INC	INC	

Table 6.15 details the frequency of each observed LB and SB allele at the D12S391 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.15: D12S391 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

D12S391									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
16	[AGAT]9[AGAC]6[AGAT]1	0.00	1.00	2	0.39	0.00	0.39	0.21	33
17	[AGAT]10[AGAC]6[AGAT]1	0.71	0.29	7	0.26	0.21	0.37	0.16	124
17	[AGAT]9[AGAC]7[AGAT]1	1.00	0.00	1	0.60	0.00	0.30	0.10	10
18	[AGAT]10[AGAC]7[AGAT]1	0.50	0.50	2	0.38	0.25	0.09	0.28	32
18	[AGAT]11[AGAC]6[AGAT]1	0.62	0.38	29	0.31	0.24	0.23	0.22	251
18	[AGAT]12[AGAC]5[AGAT]1	1.00	0.00	1	0.57	0.14	0.07	0.21	14
18.3	[AGAT]1[GAT]1[AGAT]9[AGAC]7[AGAT]1	0.50	0.50	2	0.14	0.07	0.64	0.14	14
19	[AGAT]11[AGAC]8	1.00	0.00	1	0.14	0.86	0.00	0.00	7
19	[AGAT]12[AGAC]6[AGAT]1	0.61	0.39	18	0.22	0.30	0.19	0.29	189
19	[AGAT]11[AGAC]7[AGAT]1	0.30	0.70	10	0.11	0.13	0.11	0.65	79
19.2	[AGAT]4[AT]1[AGAT]8[AGAC]6[AGAT]1	0.00	1.00	2	0.00	0.00	0.00	0.00	0
19.3	[AGAT]5[GAT]1[AGAT]7[AGAC]6[AGAT]1	0.00	1.00	1	0.20	0.00	0.40	0.40	5
19.3	[AGAT]1[GAT]1[AGAT]10[AGAC]7[AGAT]1	0.00	1.00	1	0.00	0.00	0.33	0.67	6
20	[AGAT]11[AGAC]8[AGAT]1	1.00	0.00	2	0.17	0.58	0.08	0.17	12
20	[AGAT]12[AGAC]7[AGAT]1	0.71	0.29	14	0.23	0.19	0.19	0.40	75
20	[AGAT]12[AGAC]8	1.00	0.00	1	0.46	0.00	0.46	0.08	13
20	[AGAT]13[AGAC]6[AGAT]1	0.43	0.57	7	0.21	0.33	0.13	0.33	91
20	[AGAT]14[AGAC]5[AGAT]1	1.00	0.00	1	0.67	0.33	0.00	0.00	3
21	[AGAT]12[AGAC]8[AGAT]1	1.00	0.00	5	0.13	0.61	0.17	0.09	23
21	[AGAT]13[AGAC]7[AGAT]1	0.67	0.33	3	0.12	0.18	0.36	0.33	33
21	[AGAT]13[AGAC]8	0.00	1.00	1	0.15	0.05	0.60	0.20	20
21	[AGAT]14[AGAC]6[AGAT]1	0.50	0.50	2	0.24	0.28	0.20	0.28	25
22	[AGATA]12[AGAC]10	1.00	0.00	1	0.08	0.23	0.38	0.31	13
22	[AGAT]13[AGAC]8[AGAT]1	1.00	0.00	4	0.32	0.46	0.07	0.14	28
22	[AGAT]13[AGAC]9	0.67	0.33	6	0.15	0.20	0.38	0.27	60
22	[AGAT]14[AGAC]7[AGAT]1	1.00	0.00	1	0.11	0.11	0.22	0.56	9
23	[AGAT]12[AGAC]11	1.00	0.00	1	0.00	0.00	0.00	0.00	0
23	[AGAT]13[AGAC]10	1.00	0.00	1	0.13	0.13	0.38	0.38	8
23	[AGAT]14[AGAC]9	0.33	0.67	3	0.13	0.32	0.35	0.19	31
24	[AGAT]14[AGAC]9[AGAT]1	1.00	0.00	1	0.33	0.00	0.33	0.33	3
24	[AGAT]13[AGAC]11	0.00	1.00	1	0.00	0.00	1.00	0.00	3
24	[AGAT]14[AGAC]9[AGAT]1	0.50	0.50	2	0.50	0.00	0.50	0.00	2
24	[AGAT]15[AGAC]8[AGAT]1	1.00	0.00	2	0.00	0.13	0.88	0.00	8
25	[AGAT]16[AGAC]8[AGAT]1	0.00	1.00	1	0.08	0.46	0.38	0.08	13
Total		0	53	137	295	297	311	334	1237

Table 6.16 lists each LB and SB allele present in the dataset along with read coverage for the D16S539 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Seven LB alleles and seven SB alleles are present in the dataset with no additional sequence variation present.

Table 6.16: Observed LB and SB Alleles at D16S539 Locus.

D16S539			
Sample	**Length**	**Sequence Motif**	**Reads**
JPN200	8	[GATA]8	1195
CHN007	9	[GATA]9	1004
CHN007	9	[GATA]9	1004
CHN031	9	[GATA]9	318
ILH012	9	[GATA]9	186
JPN050	9	[GATA]9	607
JPN063	9	[GATA]9	732
JPN063	9	[GATA]9	732
JPN080	9	[GATA]9	264
JPN207	9	[GATA]9	642
JPN242	9	[GATA]9	2191
JPN242	9	[GATA]9	2191
JPN275	9	[GATA]9	1214
NYAS078	9	[GATA]9	266
PHL035	9	[GATA]9	752
PHL050	9	[GATA]9	1788
PHL050	9	[GATA]9	1788
PHL055	9	[GATA]9	407
PHL071	9	[GATA]9	1512
PHL071	9	[GATA]9	1512
PHL084	9	[GATA]9	1006
PHL088	9	[GATA]9	1278
PHL097	9	[GATA]9	1937
PHL098	9	[GATA]9	2200
PHL098	9	[GATA]9	2200
PHL109	9	[GATA]9	723
PHL142	9	[GATA]9	937
PHL145	9	[GATA]9	1383
PHL154	9	[GATA]9	1264
SDHis001	9	[GATA]9	251
SDNA106	9	[GATA]9	373
SibA096	9	[GATA]9	186
SibYDy05	9	[GATA]9	1307
SibYDy05	9	[GATA]9	1307
SibYO025	9	[GATA]9	63
TXHis167	9	[GATA]9	533
ILH012	10	[GATA]10	122

ILH071	10	[GATA]10	155
JPN050	10	[GATA]10	502
JPN200	10	[GATA]10	1086
JPN274	10	[GATA]10	1140
JPN275	10	[GATA]10	967
OHHis103	10	[GATA]10	439
OHHis116	10	[GATA]10	672
PHL005	10	[GATA]10	1046
PHL035	10	[GATA]10	673
SDHis020	10	[GATA]10	237
SDNA029	10	[GATA]10	479
SDNA035	10	[GATA]10	90
SDNA052	10	[GATA]10	319
SDNA127	10	[GATA]10	298
SDNA130	10	[GATA]10	589
SibYDe54	10	[GATA]10	268
TXHis033	10	[GATA]10	541
VTAS016	10	[GATA]10	772
WANA050	10	[GATA]10	41
WANA062	10	[GATA]10	439
CHN031	11	[GATA]11	281
ILH087	11	[GATA]11	45
JPN138	11	[GATA]11	762
JPN199	11	[GATA]11	1422
JPN274	11	[GATA]11	936
PHL012	11	[GATA]11	918
PHL052	11	[GATA]11	1823
PHL052	11	[GATA]11	1823
PHL097	11	[GATA]11	1430
PHL100	11	[GATA]11	815
PHL106	11	[GATA]11	401
PHL142	11	[GATA]11	800
PHL154	11	[GATA]11	955
SDHis001	11	[GATA]11	236
SDNA003	11	[GATA]11	302
SDNA052	11	[GATA]11	232
SDNA055	11	[GATA]11	143
SDNA058	11	[GATA]11	299
SDNA127	11	[GATA]11	402
SDNA150	11	[GATA]11	114
SibA009	11	[GATA]11	807
TXHis033	11	[GATA]11	351
TXHis167	11	[GATA]11	591

WANA062	11	[GATA]11	268
ILH071	12	[GATA]12	113
ILH087	12	[GATA]12	24
ILH097	12	[GATA]12	118
ILH097	12	[GATA]12	118
JPN080	12	[GATA]12	290
JPN138	12	[GATA]12	658
JPN199	12	[GATA]12	1106
JPN260	12	[GATA]12	2204
JPN260	12	[GATA]12	2204
NYAS062	12	[GATA]12	130
NYAS062	12	[GATA]12	130
OHHis035	12	[GATA]12	901
OHHis068	12	[GATA]12	448
PHL061	12	[GATA]12	793
PHL079	12	[GATA]12	1011
PHL079	12	[GATA]12	1011
PHL088	12	[GATA]12	822
PHL100	12	[GATA]12	708
PHL110	12	[GATA]12	1092
PHL145	12	[GATA]12	997
SDHis020	12	[GATA]12	155
SDNA003	12	[GATA]12	259
SDNA035	12	[GATA]12	82
SDNA055	12	[GATA]12	242
SDNA058	12	[GATA]12	193
SDNA060	12	[GATA]12	199
SDNA126	12	[GATA]12	347
SDNA130	12	[GATA]12	487
SDNA150	12	[GATA]12	98
SibA009	12	[GATA]12	912
SibA096	12	[GATA]12	159
SibYDe54	12	[GATA]12	146
SibYO025	12	[GATA]12	14
TXHis117	12	[GATA]12	1171
TXHis117	12	[GATA]12	1171
TXHis135	12	[GATA]12	41
VTAS016	12	[GATA]12	567
WANA037	12	[GATA]12	17
WANA050	12	[GATA]12	28
WANA093	12	[GATA]12	292
WANA093	12	[GATA]12	292

JPN207	13	[GATA]13	585
NYAS078	13	[GATA]13	120
OHHis103	13	[GATA]13	269
OHHis116	13	[GATA]13	492
PHL005	13	[GATA]13	875
PHL012	13	[GATA]13	800
PHL055	13	[GATA]13	331
PHL061	13	[GATA]13	654
PHL084	13	[GATA]13	535
PHL106	13	[GATA]13	216
PHL109	13	[GATA]13	394
PHL110	13	[GATA]13	687
PHL140	13	[GATA]13	1866
PHL140	13	[GATA]13	1866
SDNA029	13	[GATA]13	467
SDNA060	13	[GATA]13	184
SDNA106	13	[GATA]13	132
SDNA126	13	[GATA]13	254
VTAS001	13	[GATA]13	1020
VTAS001	13	[GATA]13	1020
WANA037	13	[GATA]13	25
OHHis035	14	[GATA]14	670
OHHis068	14	[GATA]14	387
TXHis135	14	[GATA]14	47

Table 6.17 details the frequency of each observed LB and SB allele at the D16S539 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of

observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.17: D16S539 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

D16S539									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
8	[GATA]8	1.00	0.00	1	0.59	0.05	0.14	0.23	22
9	[GATA]9	0.89	0.11	35	0.34	0.32	0.21	0.13	241
10	[GATA]10	0.38	0.62	21	0.29	0.19	0.15	0.36	181
11	[GATA]11	0.54	0.46	24	0.25	0.20	0.28	0.27	450
12	[GATA]12	0.44	0.56	43	0.22	0.20	0.31	0.27	410
13	[GATA]13	0.67	0.33	21	0.20	0.23	0.33	0.24	219
14	[GATA]14	0.00	1.00	3	0.21	0.18	0.36	0.25	28
Total		86	62	148	400	338	419	394	1551

Table 6.18 lists each LB and SB allele present in the dataset along with read coverage for the FGA locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Fourteen LB alleles and fourteen SB alleles are present in the dataset, with no additional sequence variation present.

Table 6.18: Observed LB and SB Alleles at FGA Locus.

FGA			
Sample	**Length**	**Sequence Motif**	**Reads**
PHL035	18	[GGAA]2[GGAG]1[AAAG]10[AGAA]1[AAAA]1[GAAA]3	947
PHL140	18	[GGAA]2[GGAG]1[AAAG]10[AGAA]1[AAAA]1[GAAA]3	681
JPN275	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	1210
OHHis116	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	569
PHL012	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	463
PHL055	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	1013
PHL061	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	1149
PHL098	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	1176
PHL154	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	1236
SDHis020	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	98
SDNA055	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	192
SDNA088	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	84

SDNA126	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	139
SDNA130	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	395
TXHis033	19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	241
JPN050	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	860
JPN207	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	1204
OHHis103	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	242
PHL005	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	591
SDNA035	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	91
SDNA055	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	147
SDNA088	20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	92
CHN007	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	293
JPN080	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	156
JPN260	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	1453
PHL035	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	668
PHL052	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	954
PHL071	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	939
PHL079	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	304
PHL097	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	939
PHL106	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	418
PHL154	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	1034
SDNA127	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	133
SDNA150	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	91
SibA096	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	88
TXHis033	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	123
VTAS001	21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	508
PHL061	21.2	[GGAA]2[GGAG]1[AAAG]14[AAAA]1[AA]1[GAAA]3	904
ILH012	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	50
ILH071	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	81
JPN080	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	132
JPN199	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	435
JPN200	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	796
JPN242	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	1108
JPN260	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	1281
JPN274	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	1410
NYAS078	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	181
NYAS078	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	181
OHHis035	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	664
PHL012	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	330
PHL050	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	646
PHL052	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	732

PHL055	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	692
PHL098	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	912
PHL145	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	649
SDNA003	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	125
SDNA029	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	178
SDNA052	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	220
SDNA058	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	177
SDNA060	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	81
SDNA150	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	107
TXHis135	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	47
VTAS016	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	328
WANA062	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	182
WANA093	22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	92
NYAS062	22.2	[GGAA]2[GGAG]1[AAAG]15[AAAA]1[AA]1[GAAA]3	12
PHL142	22.2	[GGAA]2[GGAG]1[AAAG]15[AAAA]1[AA]1[GAAA]3	594
CHN031	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	207
JPN050	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	688
JPN063	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	469
JPN242	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	950
JPN274	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	1075
NYAS062	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	12
PHL071	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	754
PHL079	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	343
PHL084	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	408
PHL088	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	595
PHL100	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	747
PHL100	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	747
PHL106	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	330
PHL109	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	304
PHL110	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	788
PHL140	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	645
SDHis001	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	58
SDNA003	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	145
SDNA029	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	145
SDNA060	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	55
SDNA106	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	112
SibA009	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	374
SibYDe54	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	31
SibYDy05	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	132
SibYO025	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	26
SibYO025	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	26
TXHis135	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	38

TXHis167	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	295
WANA050	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	27
WANA050	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	27
WANA093	23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	72
ILH071	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	66
JPN063	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	420
JPN138	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	687
PHL005	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	657
PHL097	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	788
SibA096	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	91
SibYDe54	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	30
SibYDy05	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	105
TXHis167	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	184
VTAS001	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	520
VTAS016	24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	303
CHN007	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	214
CHN031	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	211
ILH012	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	73
ILH087	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	18
ILH087	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	18
JPN199	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	443
JPN200	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	632
JPN275	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	754
OHHis068	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	519
OHHis068	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	519
OHHis103	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	167
PHL050	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	486
PHL088	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	521
PHL110	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	730
PHL142	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	488
SDHis001	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	34
SDHis020	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	61
SDNA052	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	205
SDNA058	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	227
SDNA106	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	117
SDNA126	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	64
SibA009	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	334
TXHis117	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	294
WANA062	25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	196
JPN207	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	826
OHHis035	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	384

OHHis116	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	270
PHL109	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	344
PHL145	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	537
SDNA035	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	44
SDNA130	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	246
TXHis117	26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	188
JPN138	26.2	[GGAA]2[GGAG]1[AAAG]19[AAAA]1[AA]1[GAAA]3	560
ILH097	27	[GGAA]2[GGAG]1[AAAG]19[AGAA]1[AAAA]1[GAAA]3	21
ILH097	27	[GGAA]2[GGAG]1[AAAG]19[AGAA]1[AAAA]1[GAAA]3	21
SDNA127	29	[GGAA]2[GGAG]1[AAAG]21[AGAA]1[AAAA]1[GAAA]3	54
WANA037	INC		
WANA037	INC		

Table 6.19 details the frequency of each observed LB and SB allele at the FGA locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.19: FGA SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

FGA									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
18	[GGAA]2[GGAG]1[AAAG]10[AGAA]1[AAAA]1[GAAA]3	1.00	0.00	2	0.20	0.47	0.20	0.13	15
19	[GGAA]2[GGAG]1[AAAG]11[AGAA]1[AAAA]1[GAAA]3	0.46	0.54	13	0.26	0.21	0.25	0.27	107
20	[GGAA]2[GGAG]1[AAAG]12[AGAA]1[AAAA]1[GAAA]3	0.43	0.57	7	0.19	0.09	0.43	0.29	141
21	[GGAA]2[GGAG]1[AAAG]13[AGAA]1[AAAA]1[GAAA]3	0.80	0.20	15	0.22	0.22	0.37	0.19	207
21.2	[GGAA]2[GGAG]1[AAAG]14[AA]1[AAAA]1[GAAA]3	1.00	0.00	1	0.17	0.33	0.50	0.00	6
22	[GGAA]2[GGAG]1[AAAG]14[AGAA]1[AAAA]1[GAAA]3	0.56	0.44	27	0.27	0.19	0.32	0.23	270
22.2	[GGAA]2[GGAG]1[AAAG]15[AA]1[AAAA]1[GAAA]3	1.00	0.00	2	0.00	0.33	0.50	0.17	6
23	[GGAA]2[GGAG]1[AAAG]15[AGAA]1[AAAA]1[GAAA]3	0.68	0.32	31	0.28	0.31	0.19	0.22	249
24	[GGAA]2[GGAG]1[AAAG]16[AGAA]1[AAAA]1[GAAA]3	0.82	0.18	11	0.29	0.22	0.26	0.22	232
25	[GGAA]2[GGAG]1[AAAG]17[AGAA]1[AAAA]1[GAAA]3	0.42	0.58	24	0.20	0.23	0.21	0.36	168
26	[GGAA]2[GGAG]1[AAAG]18[AGAA]1[AAAA]1[GAAA]3	0.38	0.63	8	0.20	0.18	0.13	0.49	71
26.2	[GGAA]2[GGAG]1[AAAG]19[AA]1[AAAA]1[GAAA]3	1.00	0.00	1	0.00	1.00	0.00	0.00	3
27	[GGAA]2[GGAG]1[AAAG]19[AGAA]1[AAAA]1[GAAA]3	0.00	1.00	2	0.21	0.21	0.11	0.47	19
29	[GGAA]2[GGAG]1[AAAG]21[AGAA]1[AAAA]1[GAAA]3	0.00	1.00	1	0.00	0.00	0.00	0.00	0
Total		85	60	145	366	328	414	386	1494

Table 6.20 lists each LB and SB allele present in the dataset along with read coverage for the vWA locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Seven LB alleles and 19 SB alleles are present in the dataset. The "14", "15" and "20" LB alleles demonstrate two SB alleles while the "16" LB allele demonstrates three SB alleles.

Table 6.20: Observed LB and SB Alleles at vWA Locus.

vWA			
Sample	**Length**	**Sequence Motif**	**Reads**
CHN031	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	81
JPN050	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	134
JPN063	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	123
JPN138	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	268
NYAS078	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	37
PHL079	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	76
PHL106	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	93
PHL154	14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4[TAGA]1[CAGA]1[TAGA]1	171
SDNA150	14	TAGATGGA[TAGA]**10**[CAGA]**3**[TAGA]1	19
ILH097	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	15
ILH097	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	15

JPN063	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	105
JPN138	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	177
OHHis035	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	204
OHHis116	15	TAGATGGA[TAGA]**11**[CAGA]**3**[TAGA]1	88
PHL012	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	70
PHL050	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	147
PHL055	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	110
PHL084	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	51
PHL098	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	147
PHL100	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	79
SDNA029	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	72
SDNA060	15	TAGATGGA[TAGA]11[CAGA]3[TAGA]1	36
SibA096	15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	14
CHN007	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	71
ILH012	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	65
ILH012	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	65
JPN260	16	TAGATGGA[TAGA]11[CAGA]**5**	124
JPN260	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	124
JPN274	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	78
JPN274	16	TAGATGGA[TAGA]11[CAGA]5	117
JPN275	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	135
OHHis035	16	TAGATGGA[TAGA]**12**[CAGA]**3**[TAGA]1	173
OHHis116	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	110
PHL055	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	81
PHL061	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	82
PHL071	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	122
PHL100	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	69
PHL109	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	31
PHL110	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	132
PHL140	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	85
SDHis001	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	17
SDHis001	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	17
SDHis020	16	TAGATGGA[TAGA]12[CAGA]3[TAGA]1	11
SDNA003	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	41
SDNA029	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	56
SDNA058	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	59
SDNA106	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	18
SDNA126	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	91
SDNA126	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	91
SDNA127	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	30
SDNA127	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	30
SDNA130	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	111
SDNA130	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	111

SDNA150	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	12
SibYDe54	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	31
TXHis033	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	80
TXHis117	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	68
TXHis135	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	16
TXHis167	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	139
VTAS016	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	52
WANA062	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	56
WANA093	16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	26
ILH071	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	27
JPN080	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	21
JPN199	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	166
JPN199	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	166
JPN207	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	102
JPN242	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	93
JPN275	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	115
NYAS078	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	28
OHHis068	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	85
OHHis103	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	78
PHL005	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	222
PHL005	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	222
PHL035	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	65
PHL052	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	118
PHL061	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	86
PHL079	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	71
PHL084	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	38
PHL097	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	322
PHL097	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	322
PHL098	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	108
PHL106	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	43
PHL142	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	156
PHL154	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	117
SDNA052	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	57
SDNA055	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	48
SDNA055	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	48
SDNA060	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	37
SibA096	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	14
SibYDe54	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	19
SibYDy05	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	51
TXHis117	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	90
TXHis135	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	11
TXHis167	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	90
VTAS001	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	179

VTAS001	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	179
WANA062	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	48
WANA093	17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	14
CHN007	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	58
CHN031	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	59
ILH071	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	17
JPN050	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	103
JPN200	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	182
JPN200	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	182
OHHis103	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	37
PHL012	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	58
PHL050	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	136
PHL052	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	129
PHL071	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	91
PHL088	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	142
PHL110	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	118
PHL142	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	91
PHL145	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	145
PHL145	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	145
SDNA052	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	54
SDNA058	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	36
SDNA106	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	31
SibA009	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	106
SibA009	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	106
SibYDy05	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	41
TXHis033	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	61
VTAS016	18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	52
JPN080	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	17
JPN242	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	89
OHHis068	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	78
PHL035	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	55
PHL088	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	136
PHL140	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	62
SDHis020	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	13
SDNA003	19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	20
JPN207	20	TAGATGGA[TAGA]14[CAGA]5[TAGA]1	79
PHL109	20	TAGATGGA[TAGA]**15**[CAGA]**4**[TAGA]1	31
ILH087	INC		
ILH087	INC		
NYAS062	INC		
NYAS062	INC		

SDNA035	INC		
SDNA035	INC		
SDNA088	INC		
SDNA088	INC		
SibYO025	INC		
SibYO025	INC		
WANA037	INC		
WANA037	INC		
WANA050	INC		
WANA050	INC		

Table 6.21 details the frequency of each observed LB and SB allele at the vWA locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.21: vWA SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

vWA									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
14	TGGATGGA[TAGA]3[TGGA]1[TAGA]3[CAGA]4 [TAGA]1[CAGA]1[TAGA]1	1.00	0.00	8	0.06	0.57	0.19	0.18	146
14	TAGATGGA[TAGA]10[CAGA]3[TAGA]1	0.00	1.00	1	0.11	0.00	0.78	0.11	9
15	TAGATGGA[TAGA]10[CAGA]4[TAGA]1	0.75	0.25	12	0.64	0.04	0.17	0.14	92
15	TAGATGGA[TAGA]11[CAGA]3[TAGA]1	0.00	1.00	1	0.35	0.05	0.42	0.18	65
16	TAGATGGA[TAGA]11[CAGA]5	1.00	0.00	2	0.00	0.00	0.00	0.00	0
16	TAGATGGA[TAGA]11[CAGA]4[TAGA]1	0.37	0.63	35	0.23	0.15	0.23	0.38	313
16	TAGATGGA[TAGA]12[CAGA]3[TAGA]1	0.00	1.00	2	0.54	0.03	0.30	0.13	70
17	TAGATGGA[TAGA]12[CAGA]4[TAGA]1	0.68	0.32	37	0.16	0.24	0.30	0.30	359

18	TAGATGGA[TAGA]13[CAGA]4[TAGA]1	0.75	0.25	24	0.20	0.25	0.32	0.23	265
19	TAGATGGA[TAGA]14[CAGA]4[TAGA]1	0.63	0.38	8	0.19	0.27	0.31	0.23	102
20	TAGATGGA[TAGA]14[CAGA]5[TAGA]1	1.00	0.00	1	0.25	0.25	0.00	0.50	4
20	TAGATGGA[TAGA]15[CAGA]4[TAGA]1	1.00	0.00	1	0.04	0.40	0.24	0.32	25
	Total	82	50	132	336	332	401	381	1450

Table 6.22 lists each LB and SB allele present in the dataset along with read coverage for the DYS390 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Six LB alleles and nine SB alleles are present in the dataset. The "25" LB allele demonstrates two SB alleles while the "21" LB allele demonstrates three SB alleles.

Table 6.22: Observed LB and SB Alleles at DYS390 Locus.

DYS390				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
JPN200	AS	21	[TAGA]**13**[CAGA]**8**	1453
JPN274	AS	21	[TAGA]4[CAGA]1[TAGA]**8**[CAGA]**8**	1947
PHL084	AS	21	[TAGA]4[CAGA]1[TAGA]**9**[CAGA]**7**	681
JPN080	AS	22	[TAGA]4[CAGA]1[TAGA]9[CAGA]8	341
OHHis103	NA	22	[TAGA]4[CAGA]1[TAGA]9[CAGA]8	309
SDNA106	NA	22	[TAGA]4[CAGA]1[TAGA]9[CAGA]8	391
JPN260	AS	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	1937
OHHis116	NA	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	597
PHL106	AS	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	728
PHL109	AS	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	891
PHL110	AS	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	1263
SibYDy05	AS	23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	282
JPN050	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	24
JPN138	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	252
OHHis035	NA	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	135
OHHis068	NA	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	443
PHL035	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	660
PHL055	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	1209
PHL061	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	1597
PHL079	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	528
PHL145	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	1040
SDHis001	NA	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	126
SDNA126	NA	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	287
SDNA130	NA	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	688
VTAS001	AS	24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	814
JPN242	AS	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	2001
PHL100	AS	25	[TAGA]4[CAGA]1[TAGA]**12**[CAGA]**8**	898
PHL140	AS	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	944

PHL142	AS	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	765
PHL154	AS	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	1483
SDHis020	NA	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	176
SDNA127	NA	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	115
SDNA150	NA	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	184
VTAS016	AS	25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	656
ILH071	NA	26	[TAGA]4[CAGA]1[TAGA]13[CAGA]8	76

Table 6.23 details the frequency of each observed LB and SB allele at the DYS390 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.23: DYS390 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

DYS390									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
21	[TAGA]13[CAGA]8	1.00	0.00	1	0.00	0.00	0.00	0.00	0.00%
21	[TAGA]4[CAGA]1[TAGA]8[CAGA]8	1.00	0.00	1	0.95	0.00	0.05	0.00	22
21	[TAGA]4[CAGA]1[TAGA]9[CAGA]7	1.00	0.00	1	0.00	1.00	0.00	0.00	1
22	[TAGA]4[CAGA]1[TAGA]9[CAGA]8	0.33	0.67	3	0.21	0.11	0.58	0.11	19
23	[TAGA]4[CAGA]1[TAGA]10[CAGA]8	0.63	0.38	8	0.07	0.53	0.33	0.06	81
24	[TAGA]4[CAGA]1[TAGA]11[CAGA]8	0.62	0.38	13	0.06	0.33	0.47	0.14	95
25	[TAGA]4[CAGA]1[TAGA]11[CAGA]9	0.63	0.38	8	0.00	1.00	0.00	0.00	3
25	[TAGA]4[CAGA]1[TAGA]12[CAGA]8	0.50	0.50	2	0.03	0.56	0.38	0.03	32
26	[TAGA]4[CAGA]1[TAGA]13[CAGA]8	0.00	1.00	1	0.00	0.00	1.00	0.00	2
Total		23	15	38	38	98	98	21	255

Table 6.24 lists each LB and SB allele present in the dataset along with read coverage for the DYS392 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Five LB alleles and five SB alleles are present in the dataset with no additional sequence variation.

Table 6.24: Observed LB and SB Alleles at DYS392 Locus.

DYS392				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
JPN242	AS	11	[ATA]11	1611
JPN260	AS	11	[ATA]11	839
OHHis103	NA	11	[ATA]11	785
OHHis116	NA	11	[ATA]11	1534
PHL084	AS	11	[ATA]11	296
SDHis020	NA	11	[ATA]11	453
SDNA150	NA	11	[ATA]11	98
PHL142	AS	12	[ATA]12	668
SDNA106	NA	12	[ATA]12	160
JPN080	AS	13	[ATA]13	47
PHL035	AS	13	[ATA]13	105
PHL055	AS	13	[ATA]13	170
PHL061	AS	13	[ATA]13	154
PHL079	AS	13	[ATA]13	89
PHL100	AS	13	[ATA]13	246
PHL140	AS	13	[ATA]13	564
PHL145	AS	13	[ATA]13	559
PHL154	AS	13	[ATA]13	551
JPN050	AS	14	[ATA]14	35
JPN200	AS	14	[ATA]14	104
PHL106	AS	14	[ATA]14	37
PHL109	AS	14	[ATA]14	460
PHL110	AS	14	[ATA]14	237
SDNA127	NA	14	[ATA]14	25
SDNA130	NA	14	[ATA]14	70
VTAS001	AS	14	[ATA]14	54
VTAS016	AS	14	[ATA]14	23
JPN274	AS	15	[ATA]15	11
OHHis068	NA	15	[ATA]15	15
ILH071	NA	INC		
JPN138	AS	INC		

OHHis035	NA	INC		
SDHis001	NA	INC		
SDNA126	NA	INC		
SibYDy05	AS	INC		

Table 6.25 details the frequency of each observed LB and SB allele at the DYS392 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.25: DYS392 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

DYS392									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
11	[ATA]11	0.43	0.57	7	0.03	0.13	0.46	0.06	85
12	[ATA]12	0.50	0.50	2	0.00	0.79	0.14	0.07	14
13	[ATA]13	1.00	0.00	9	0.09	0.37	0.44	0.09	106
14	[ATA]14	0.78	0.22	9	0.00	0.86	0.07	0.07	44
15	[ATA]15	0.50	0.50	2	0.00	0.67	0.00	0.33	3
Total		21	8	29	40	101	91	20	252

Table 6.26 lists each LB and SB allele present in the dataset along with read coverage for the DYS438 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Five LB alleles and eight SB

alleles are present in the dataset. The “10”, “11” and “12” LB alleles demonstrate two SB alleles.

Table 6.26: Observed LB and SB Alleles at DYS438 Locus.

DYS438			
Sample	Length	Sequence Allele	Reads
OHHis116	9	[TTTTC]9	191
JPN242	10	[TTTTC]10	2841
JPN260	10	[TTTTC]10	4175
JPN274	10	[TTTTC]10	3613
OHHis103	10	[TTTTC]10	24
PHL035	10	[TTTTC]10	2908
PHL055	10	[TTTTC]10	1655
PHL061	10	[TTTTC]10	3829
PHL079	10	[TTTTC]10	1185
PHL084	10	[TTTTC]10	1603
PHL100	10	[TTTTC]10	806
PHL106	10	[TTTTC]10	646
PHL140	10	[TTTTC]10	1921
PHL142	10	[TTTTC]10	732
PHL154	10	[TTTTC]10	1261
SDHis020	10	[TTTTC]10	79
SDNA106	10	[TTTTC]9[TTTTT]1	35
SDNA106	10	[TTTTC]10	151
SDNA150	10	[TTTTC]10	107
ILH071	11	[TTTTC]9[TTTTT]1[TTTTC]1	39
ILH071	11	[TTTTC]11	121
JPN050	11	[TTTTC]11	2520
JPN200	11	[TTTTC]11	2468
OHHis035	11	[TTTTC]9[TTTTT]1[TTTTC]1	25
OHHis035	11	[TTTTC]11	29
OHHis068	11	[TTTTC]11	17
PHL109	11	[TTTTC]9[TTTTT]1[TTTTC]1	120
PHL109	11	[TTTTC]11	362
PHL110	11	[TTTTC]9[TTTTT]1[TTTTC]1	151
PHL110	11	[TTTTC]11	444
SDHis001	11	[TTTTC]11	48
SDNA126	11	[TTTTC]11	51
SDNA127	11	[TTTTC]11	40
SDNA130	11	[TTTTC]9[TTTTT]1[TTTTC]1	53

SDNA130	11	[TTTTC]11	163
SibYDy05	11	[TTTTC]11	487
VTAS001	11	[TTTTC]9[TTTTT]1[TTTTC]1	131
VTAS001	11	[TTTTC]11	406
VTAS016	11	[TTTTC]9[TTTTT]1[TTTTC]1	95
VTAS016	11	[TTTTC]11	273
PHL145	12	[TTTTC]9[TTTTT]1[TTTTC]2	59
PHL145	12	[TTTTC]12	173
JPN080	13	[TTTTC]13	274
JPN138	13	[TTTTC]13	980

Table 6.27 details the frequency of each observed LB and SB allele at the DYS438 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group.

Table 6.27: DYS438 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

DYS438									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
9	[TTTTC]9	0.00	1.00	1	0.17	0.17	0.42	0.25	12
10	[TTTTC]10	0.76	0.24	17	0.06	0.64	0.26	0.04	103
10	[TTTTC]9[TTTTT]1	0.00	1.00	1	0.00	0.00	0.00	0.00	0
11	[TTTTC]11	0.50	0.50	14	0.28	0.47	0.19	0.05	78
11	[TTTTC]9[TTTT]1[TTTTC]1	0.57	0.43	7	0.00	0.00	0.00	0.00	0
12	[TTTTC]9[TTTTT]1[TTTTC]2	1.00	0.00	1	0.00	0.00	0.00	0.00	0
12	[TTTTC]12	1.00	0.00	1	0.14	0.00	0.73	0.14	74
13	[TTTTC]13	1.00	0.00	2	0.00	0.25	0.75	0.00	4
Total		28	16	44	40	106	105	21	271

Table 6.28 lists each LB and SB allele present in the dataset along with read coverage for the DYS448 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Four LB alleles and six SB alleles are present in the dataset. The "19" and "20" LB alleles demonstrate two SB alleles.

Table 6.28 Observed LB and SB Alleles at DYS448 Locus.

DYS448				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
JPN242	AS	17	[AGAGAT]**10**[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]**7**	695
PHL061	AS	17	[AGAGAT]**10**[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]**7**	464
JPN080	AS	18	[AGAGAT]10[ATAGAG]2[AGATA]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	71
JPN138	AS	18	[AGAGAT]10[ATAGAG]2[AGATA]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	84
PHL106	AS	18	[AGAGAT]10[ATAGAG]2[AGATA]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	34
PHL140	AS	18	[AGAGAT]10[ATAGAG]2[AGATA]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	110
JPN260	AS	19	[AGAGAT]**10**[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]**9**	339
JPN274	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	297
PHL035	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	174
PHL055	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	276
PHL079	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	29
PHL142	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	40
PHL145	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	65
PHL154	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	110
SDNA130	NA	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	33
SibYDy05	AS	19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]8	32

JPN050	AS	20	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]9	117
JPN200	AS	20	[AGAGAT]**12**[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]**8**	207
PHL110	AS	20	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]9	45
VTAS001	AS	20	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1[AGAGAT]9	45
ILH071	NA	INC		
OHHis035	NA	INC		
OHHis068	NA	INC		
OHHis103	NA	INC		
OHHis116	NA	INC		
PHL084	AS	INC		
PHL100	AS	INC		
PHL109	AS	INC		
SDHis001	NA	INC		
SDHi020	NA	INC		
SDNA106	NA	INC		
SDNA126	NA	INC		
SDNA127	NA	INC		
SDNA150	NA	INC		
VTAS016	AS	INC		

Table 6.29 details the frequency of each observed LB and SB allele at the DYS448 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples used for each group. NA samples at this locus experienced high allelic dropout and were unable to successfully type more than one sample with sufficient read coverage.

Table 6.29: DYS448 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

DYS448									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
17	[AGAGAT]10[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1 [AGAGAT]7	1.00	0.00	2	0.00	0.00	0.00	0.00	0
18	[AGAGAT]10[ATAGAG]2[AGATA]3[ATAGAT]1[AGAGAA]1 [AGAGAT]8	1.00	0.00	4	0.00	0.91	0.09	0.00	32
19	[AGAGAT]10[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1 [AGAGAT]9	1.00	0.00	1	0.00	0.50	0.50	0.00	2
19	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1 [AGAGAT]8	0.89	0.11	9	0.11	0.26	0.53	0.10	102
20	[AGAGAT]11[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1 [AGAGAT]9	1.00	0.00	3	0.00	0.86	0.07	0.07	28
20	[AGAGAT]12[ATAGAG]2[AGATAG]3[ATAGAT]1[AGAGAA]1 [AGAGAT]8	1.00	0.00	1	0.16	0.16	0.62	0.06	50
	Total	19	1	20	19	89	91	15	214

Table 6.30 lists each LB and SB allele present in the dataset along with read coverage for the DYS635 locus. Length-based alleles with different sequence motifs are bolded the first time they appear in the table to better display sequence differences. Seven LB alleles and eight SB alleles are present in the dataset. The "22" LB allele demonstrate two SB alleles.

Table 6.30: Observed LB and SB Alleles at DYS635 Locus.

DYS635				
Sample	**Pop**	**Length**	**Sequence Motif**	**Reads**
PHL142	AS	19	[TAGA]10[TACA]1[TAGA]2[TACA]2[TAGA]4	98
JPN080	AS	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	101
PHL109	AS	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	120
PHL110	AS	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	175
SDNA106	NA	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	24
VTAS001	AS	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	130
VTAS016	AS	20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	76
JPN050	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	587
JPN138	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	394
JPN260	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	846
JPN274	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	685
PHL061	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	993
PHL079	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	97
PHL106	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	68
PHL140	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	190

PHL154	AS	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	146
SDHis020	NA	21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	23
JPN200	AS	22	[TAGA]8[TACA]2[TAGA]2[TACA]2[TAGA]2[TACA]2[TAGA]4	582
PHL035	AS	22	[TAGA]12[TACA]2[TAGA]2[TACA]2[TAGA]4	714
PHL055	AS	22	[TAGA]12[TACA]2[TAGA]2[TACA]2[TAGA]4	454
SDNA130	NA	22	[TAGA]8[TACA]2[TAGA]2[TACA]2[TAGA]2[TACA]2[TAGA]4	51
SibYDy05	AS	22	[TAGA]12[TACA]2[TAGA]2[TACA]2[TAGA]4	127
PHL145	AS	23	[TAGA]9[TACA]2[TAGA]2[TACA]2[TAGA]2[TACA]2[TAGA]4	95
PHL084	AS	24	[TAGA]14[TACA]2[TAGA]2[TACA]2[TAGA]4	90
JPN242	AS	25	[TAGA]15[TACA]2[TAGA]2[TACA]2[TAGA]4	550
PHL100	AS	25	[TAGA]15[TACA]2[TAGA]2[TACA]2[TAGA]4	51
ILH071	NA	INC		
OHHis035	NA	INC		
OHHis068	NA	INC		
OHHis103	NA	INC		
OHHis116	NA	INC		
SDHis001	NA	INC		
SDNA126	NA	INC		
SDNA127	NA	INC		
SDNA150	NA	INC		

Table 6.31 details the frequency of each observed LB and SB allele at the DYS635 locus for the NA and AS groups in this dataset as well as those compiled in Novroski et al. (2016). The "Total" column for the NA and AS groups represents the total number of SB alleles observed in this dataset. The NA and AS frequencies were calculated from the number of occurrences in each group (NA or AS) compared to the total number of observations among both groups. The "Total" column for the AFA, ASN, CAU, and HIS groups represents the total number of SB alleles observed in the current literature. The frequencies for these four groups was calculated from the number of occurrences in each group compared to the total number of observations, among all four groups. The "Total" row represents the total number of samples

used for each group. NA samples at this locus experienced high allelic dropout and were unable to successfully type more than three sample with sufficient read coverage.

Table 6.31: DYS635 SB Alleles Present in the Dataset with Observed Frequencies per U.S. Population Group.

DYS635									
LB	SB	AS	NA	Total	AFA	ASN	CAU	HIS	Total
19	[TAGA]10[TACA]1[TAGA]2[TACA]2[TAGA]4	1.00	0.00	1	0.00	1.00	0.00	0.00	2
20	[TAGA]10[TACA]2[TAGA]2[TACA]2[TAGA]4	0.83	0.17	6	0.02	0.92	0.06	0.00	49
21	[TAGA]11[TACA]2[TAGA]2[TACA]2[TAGA]4	0.90	0.10	10	0.25	0.48	0.18	0.08	60
22	[TAGA]8[TACA]2[TAGA]2[TACA]2[TAGA]2[TACA]2[TAGA]4	0.50	0.50	2	0.00	0.00	0.40	0.60	5
22	[TAGA]12[TACA]2[TAGA]2[TACA]2[TAGA]4	1.00	0.00	3	0.25	0.33	0.39	0.03	36
23	[TAGA]9[TACA]2[TAGA]2[TACA]2[TAGA]2[TACA]2[TAGA]4	1.00	0.00	1	0.10	0.00	0.75	0.15	60
24	[TAGA]14[TACA]2[TAGA]2[TACA]2[TAGA]4	1.00	0.00	1	0.00	0.80	0.20	0.00	5
25	[TAGA]15[TACA]2[TAGA]2[TACA]2[TAGA]4	1.00	0.00	2	0.33	0.33	0.33	0.00	3
	Total	23	3	26	32	93	77	18	220

Sequence-Based Allele Distributions

Table 6.32 displays the Hardy-Weinberg equilibrium test and significance values for the D2S441 locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. LB alleles for both NA and AS groups and SB alleles for the NA group were non-significant, however, the SB alleles of the AS group significantly deviated from HWE.

Table 6.32: HWE Test at D2S441 Locus with LB and SB Alleles for AS and NA Samples Separately.

D2S441				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.76744	0.79289	0.38086	0.00039
SB	0.7907	0.82408	0.01951	0.00015
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.58065	0.69646	0.05898	0.0002
SB	0.74194	0.82813	0.15318	0.00023

Table 6.33 displays the Hardy-Weinberg equilibrium test and significance values for the D2S1338 locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. LB alleles for both NA and AS groups and SB alleles for the NA group were non-significant, however, the SB alleles of the AS group significantly deviated from HWE. LB alleles for both groups and SB alleles for the AS group were non-significant, but the SB alleles for the NA group significantly deviated from HWE.

Table 6.33: HWE Test at D2S1338 Locus with LB and SB Alleles for AS and NA Samples Separately.

D2S1338				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.93023	0.87168	0.75926	0.00035
SB	0.93023	0.93187	0.12478	0.00016
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.87879	0.88951	0.30829	0.00036
SB	0.87879	0.91469	0.01557	0.00007

Table 6.34 displays the Hardy-Weinberg equilibrium test and significance values for the D7S820 locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. Both LB and SB alleles for both groups were non-significant.

Table 6.34: HWE Test at D7S820 Locus with LB and SB Alleles for AS and NA Samples Separately.

D7S820				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.74419	0.75568	0.25989	0.00038
SB	0.74419	0.76088	0.24504	0.00042
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.6129	0.70598	0.77374	0.00039
SB	0.6129	0.70598	0.76297	0.00034

Table 6.35 displays the Hardy-Weinberg equilibrium test and significance values for the D12S391 locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. Both LB and SB alleles for both groups were non-significant.

Table 6.35: HWE Test at D12S391 Locus with LB and SB Alleles for AS and NA Samples Separately.

D12S391				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.72727	0.86155	0.11422	0.00023
SB	0.86364	0.91588	0.52895	0.00032
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.95833	0.84131	0.45655	0.00029
SB	0.95833	0.90514	0.81146	0.00022

Table 6.36 displays the Hardy-Weinberg equilibrium test and significance values for the D16S539 locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. Both LB and SB alleles for both groups were non-significant.

Table 6.36: HWE Test at D16S539 Locus with LB and SB Alleles for AS and NA Samples Separately.

D16S539				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.66667	0.76783	0.6709	0.00045
SB	0.67647	0.77173	0.73107	0.00039
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.87179	0.8012	0.47448	0.00046
SB	0.87179	0.8012	0.47146	0.00046

Table 6.37 displays the Hardy-Weinberg equilibrium test and significance values for the FGA locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. Both LB and SB alleles for both groups were non-significant.

Table 6.37: HWE Test at FGA Locus with LB and SB Alleles for AS and NA Samples Separately.

FGA				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.91111	0.85793	0.72587	0.00041
SB	0.91111	0.86517	0.67296	0.00033
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.88889	0.87002	0.17072	0.00032
SB	0.88889	0.87002	0.16967	0.00032

Table 6.38 displays the Hardy-Weinberg equilibrium test and significance values for the vWA locus. HWE tests for both LB and SB alleles are displayed separately for AS and NA samples. Both LB and SB alleles for both groups were non-significant.

Table 6.38: HWE Test at vWA Locus with LB and SB Alleles for AS and NA Samples Separately.

vWA				
Asian Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.78049	0.80939	0.91197	0.00034
SB	0.82927	0.81451	0.26287	0.00051
Native American Samples				
Allele	Obs. Het	Exp. Het	P-value	St. Dev.
LB	0.73077	0.71719	0.95906	0.00017
SB	0.73077	0.75038	0.6263	0.00048

Fisher's Exact Tests

Table 6.39 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the D2S441 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significantly differentfor both LB and SB alleles at this locus.

Table 6.39: Fisher's Exact Test Between AS and NA Samples for D2S441 Locus for Both LB and SB Alleles.

D2S441: Length					D2S441: Sequence			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	22.537	8	0.001		Pearson Chi-Square	25.888	11	0.002
Likelihood Ratio	26.458	8	0.001		Likelihood Ratio	30.212	11	0.002
Fisher's Exact Test	22.42		0.001		Fisher's Exact Test	25.764		0.002
N Valid Cases	148				N Valid Cases	148		

Table 6.40 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the D2S1338 locus. This test was performed for LB and SB alleles separately.

The SB allelic frequency distribution at the D2S1338 locus was significant but was not significant for the LB allelic distribution.

Table 6.40: Fisher's Exact Test Between AS and NA Samples for D2S1338 Locus for Both LB and SB Alleles.

D2S1338 Length					D2S1338 Sequence			
	Value	df	Exact Significance (2-sided)			Value	df	Exact Significance (2-sided)
Pearson Chi-Square	10.838	10	0.375		Pearson Chi-Square	46.764	26	0.001
Likelihood Ratio	11.597	10	0.403		Likelihood Ratio	60.998	26	0.001
Fisher's Exact Test	10.647		0.371		Fisher's Exact Test	43.673		0.001
N Valid Cases	146				N Valid Cases			

Table 6.41 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the D7S820 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significant for both LB and SB alleles at this locus.

Table 6.41: Fisher's Exact Test Between AS and NA Samples for D7S820 Locus for Both LB and SB Alleles.

D7S820 Length					D7S820 Sequence			
	Value	df	Exact Significance (2-sided)			Value	df	Exact Significance (2-sided)
Pearson Chi-Square	21.267	7	0.001		Pearson Chi-Square	21.89	8	0.001
Likelihood Ratio	22.499	7	0.002		Likelihood Ratio	23.485	8	0.002
Fisher's Exact Test	20.498		0.001		Fisher's Exact Test	21.066		0.002
N Valid Cases	148				N Valid Cases	148		

Table 6.42 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the D12S391 locus. This test was performed for LB and SB alleles separately.

The frequency distributions between the two groups were not statistically significant for both LB and SB alleles at this locus.

Table 6.42: Fisher's Exact Test Between AS and NA Samples for D12S391 Locus for Both LB and SB Alleles.

D12S391 Length					D12S391 Sequence			
	Value	df	Exact Significance (2-sided)			Value	df	Exact Significance (2-sided)
Pearson Chi-Square	15.603	12	0.194		Pearson Chi-Square	35.628	32	0.247
Likelihood Ratio	17.908	12	0.228		Likelihood Ratio	46.039	32	0.218
Fisher's Exact Test	14.568		0.203		Fisher's Exact Test	33.58		0.233
N Valid Cases	136				N Valid Cases	136		

Table 6.43 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the D16S539 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were not statistically significant for both LB and SB alleles at this locus.

Table 6.43: Fisher's Exact Test Between AS and NA Samples for D16S539 Locus for Both LB and SB Alleles.

D16S539 Length					D16S539 Sequence			
	Value	df	Exact Significance (2-sided)			Value	df	Exact Significance (2-sided)
Pearson Chi-Square	19.339	15	0.178		Pearson Chi-Square	19.339	15	0.178
Likelihood Ratio	22.133	15	0.246		Likelihood Ratio	22.133	15	0.256
Fisher's Exact Test	18.509		0.181		Fisher's Exact Test	18.509		0.181
N Valid Cases	73				N Valid Cases	73		

Table 6.44 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the FGA locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significant for both LB and SB alleles at this locus.

Table 6.44: Fisher's Exact Test Between AS and NA Samples for FGA Locus for Both LB and SB Alleles.

FGA Length					**FGA Sequence**			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	20.785	13	0.045		Pearson Chi-Square	20.785	13	0.045
Likelihood Ratio	24.464	13	0.044		Likelihood Ratio	24.464	13	0.044
Fisher's Exact Test	19.246		0.06		Fisher's Exact Test	19.246		0.06
N Valid Cases	145				N Valid Cases	145		

Table 6.45 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the vWA locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significant for both LB and SB alleles at this locus.

Table 6.45: Fisher's Exact Test Between AS and NA Samples for vWA Locus for Both LB and SB Alleles.

vWA Length				**vWA Sequence**			
Length	Value	df	Exact Significance (2-sided)	Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	15.236	6	0.014	Pearson Chi-Square	25.84	10	0.001
Likelihood Ratio	16.347	6	0.017	Likelihood Ratio	30.765	10	0.001
Fisher's Exact Test	14.378		0.017	Fisher's Exact Test	24.592		0.001
N Valid Cases	134			N Valid Cases	134		

Table 6.46 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the DYS390 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were not statistically significant for both LB and SB alleles at this locus.

Table 6.46: Fisher's Exact Test Between AS and NA Samples for DYS390 Locus for Both LB and SB Alleles.

DYS390 Length					DYS390 Sequence			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	5.809	5	0.331		Pearson Chi-Square	4.145	7	0.912
Likelihood Ratio	6.998	5	0.338		Likelihood Ratio	5.097	7	0.905
Fisher's Exact Test	5.188		0.38		Fisher's Exact Test	4.441		0.882
N Valid Cases	35				N Valid Cases	35		

Table 6.47 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the DYS392 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significant for both LB and SB alleles at this locus.

Table 6.47: Fisher's Exact Test Between AS and NA Samples for DYS392 Locus for Both LB and SB Alleles.

DYS392 Length				DYS392 Sequence			
Length	Value	df	Exact Significance (2-sided)	Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	7.625	4	0.108	Pearson Chi-Square	7.625	4	0.108
Likelihood Ratio	9.522	4	0.085	Likelihood Ratio	9.522	4	0.085
Fisher's Exact Test	8.226		0.047	Fisher's Exact Test	8.226		0.047
N Valid Cases	29			N Valid Cases	29		

Table 6.48 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the DYS438 locus. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were not statistically significant for both LB and SB alleles at this locus.

Table 6.48: Fisher's Exact Test Between AS and NA Samples for DYS438 Locus for Both LB and SB Alleles.

DYS438 Length					**DYS438 Sequence**			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	5.759	4	0.25		Pearson Chi-Square	8.248	7	85
Likelihood Ratio	7.348	4	0.164		Likelihood Ratio	10.163	7	0.283
Fisher's Exact Test	4.819		0.25		Fisher's Exact Test	7.622		0.291
N Valid Cases	44				N Valid Cases	44		

Table 6.49 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the DYS448 locus. As discussed earlier, this locus suffered from high allelic dropout for many of the NA samples. To address this, the Hispanic (HIS) samples from Novroksi et al (2016) were used in conjunction with the few NA samples that did amplify. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were not statistically significant for both LB and SB alleles at this locus.

Table 6.49: Fisher's Exact Test Between AS and NA/HIS Samples for DYS448 Locus for Both LB and SB Alleles.

DYS448 Length					DYS448 Sequence			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	7.261	10	0.565		Pearson Chi-Square	15.9	18	0.478
Likelihood Ratio	9.428	10	0.454		Likelihood Ratio	20.96	18	0.129
Fisher's Exact Test	11.514		0.595		Fisher's Exact Test	25.888		0.224
N Valid Cases	41				N Valid Cases	41		

Table 6.50 displays Fisher's exact test for differences in allele distributions between AS and NA groups at the DYS635 locus. As discussed earlier, this locus suffered from high allelic dropout for many of the NA samples. To address this, the Hispanic (HIS) samples from Novroksi et al (2016) were used in conjunction with the few NA samples that did amplify. This test was performed for LB and SB alleles separately. The frequency distributions between the two groups were statistically significant for both LB and SB alleles at this locus.

Table 6.50: Fisher's Exact Test Between AS and NA/HIS Samples for DYS635 Locus for Both LB and SB Alleles.

DYS635 Length					DYS635 Sequence			
Length	Value	df	Exact Significance (2-sided)		Length	Value	df	Exact Significance (2-sided)
Pearson Chi-Square	14.03	6	0.013		Pearson Chi-Square	19.454	10	0.007
Likelihood Ratio	16.469	6	0.015		Likelihood Ratio	23.533	10	0.011
Fisher's Exact Test	14.226		0.009		Fisher's Exact Test	18.454		0.007
N Valid Cases	47				N Valid Cases	47		

Discussion

Allele Frequencies

To ensure the observed isoalleles and repeat pattern variants were not the result of sequencing error, the sequencing error rate was taken into account. The Illumina MiSeq FGx platform has sequencing error rate of 10^{-2} to 10^{-3} (1 nucleotide in 100-1,000 bases) (Kircher and Kelso, 2010; Fox et al., 2014; Sharma et al., 2017). The most common errors are represented by single nucleotide substitutions resulting from errors during amplification and sequencing due to polymerase mistakes and incorrect base identification by the analysis software. To avoid classifying alleles due to sequencing error, adequate coverage is required for each locus, particularly for any assumed isoalleles or repeat pattern variants.

The D12S391 locus displayed the highest level of sequence-based variance with an additional 19 alleles observed. The D2S1338 and vWA loci displayed the second and third highest levels of SB allele variance, respectively. The D2S441, and D7S820 loci displayed lower levels of SB variance while the FGA and D16S539 loci displayed no increase in SB alleles. These results mirror recently documented allele counts with the exception of the FGA and D16S539 loci (Planz et al., 2012; Scheible et al., 2014, Gelardi et al., 2014, Zeng et al., 2015, Gettings et al., 2016, Novroski et al., 2016; van der Gaag et al., 2016, Zhao et al., 2016). Scheible et al. (2014), Gettings et al (2016), van der Gaag et al. (2016) and Novroski et al. (2016) all reported an increase in SB alleles at these loci. This is likely due to all of these works having higher sample sizes and more diverse populations than the present study.

For the Y-STRs, the DYS390 and DYS438 loci had the highest SB allele variance while DYS448 and DYS635 had a lower level of SB allelic variance. This is similar to recently reported levels of variance at Y-STR loci (D'Amato et al., 2010; Zhao et al., 2015, Kwon et al.,

2016; Novroksi et al., 2016, Wendt et al., 2016, Just et al., 2017). While Kwon et al (2016) and Novroski et al (2016) found SB allele variance at the DYS392 locus, the present data displayed no additional SB alleles, likely resulting from fewer samples for Y-STR analyses here.

The LB and SB alleles of the D7S820, D12S391, D16S539, FGA, and vWA loci were within Hardy-Weinberg equilibrium with non-significant values. However, the D2S441 locus deviated from HWE for the SB alleles for the AS group while the D2S1338 locus deviated from HWE for SB alleles for the NA group. The SB alleles at these two loci exhibit higher than expected levels of homozygosity. This might be the result of deviation from the assumptions of the Hardy Weinberg test or possibly due to null alleles. Null alleles may result from allele dropout that present as homozygotes in the data set, particularly when using kits with more than one primer set like that used for this study.

Isoallele Distributions

A Fisher's exact test was performed to statistically analyze the frequency distributions of LB and SB alleles for the NA and AS groups. The following loci all produced significant values: D2S441 (LB p-value: 0.001; SB p-value: 0.002), D7S820 (LB p-value: 0.001; SB p-value: 0.002), vWA (LB p-value: 0.017; SB p-value: 0.001), DYS392 (LB p-value: 0.047; SB p-value: 0.047), and DYS635 loci (LB p-value: 0.009; SB p-value: 0.007). These loci demonstrate that both LB and SB alleles are informative in population genetic studies as they exhibit different frequency distributions between the NA and AS groups. However, the DYS392, D2S441, and D7S820 loci produce the same statistical significance whether LB alleles or SB alleles are examined. This suggests that this locus would not provide addition genetic characterization of populations if SB alleles were used in conjunction with LB alleles. The vWA and DYS635 loci demonstrate higher statistical significance comparing the SB alleles to the LB alleles. This

indicates these loci would be good candidates for NGS in anthropological genetics as they provide increased genetic characterization beyond LB alleles.

At the D2S1338 locus, the Native American and Asian population groups were not statistically significantly different when only using length-based alleles. However, when using sequence-based alleles, the two groups are statistically significantly different (0.00012). This locus would be an ideal candidate for anthropological studies seeking better markers using NGS for genetic characterization of populations as the LB alleles provide no significant population differentiation, but the SB alleles do illuminate population differentiation. The FGA locus was nearly significant (0.06) for the LB and SB alleles. The D16S539, D12S391, DYS390, DYS438, and DYS448 loci demonstrated no significant differences between the LB and SB frequency distributions of the NA and AS groups. This suggests that these five loci would not be substantially informative for population genetic studies seeking to expand genetic characterization of populations.

Chapter 7: Conclusions

MtDNA analyses in anthropological research primarily focus on variants within the hypervariable segments or the control region to make haplogroup assessments. These regions are more informative for haplogroup assignments than the coding region. However, variants found outside the control region are becoming more commonplace for the reporting of population genetic variation. While hypervariable segment and control region sequencing using Sanger sequencing requires two to three primer sets followed by two to three PCR cycles, whole mitogenome Sanger sequencing is very burdensome as it requires numerous primer sets and PCR reactions and requires a greater amount of the sample being tested. As genetic technologies like NGS are progressing and becoming more utilized by various research disciplines, anthropological research is slowly adopting them as well. NGS provides a manner of sequencing the whole mitogenome using only two primer pairs and two PCR cycles. The added data provided by whole mitogenome sequencing can provide increased information for understanding genetic variation within a population, for tracing specific haplogroups through migratory processes over time, and for identifying subtypes within haplogroups not previously recognized. Further, mitogenome sequencing using NGS methods can substantially improve our understanding of genetic diversity and our ability to accurately assign haplogroups.

As demonstrated here, haplogrouping was accurate for 95% of samples when using less than the full mitogenome. Using the control region alone, 50% of samples were precisely haplogrouped and 82% of Native American haplogroups were distinguishable from Asian haplogroups. While many samples were precisely haplogrouped with accurate ancestry predictions using control region data, at least one in six samples did not contain variants within the control region capable of distinguishing Native American haplogroups from Asian B4

haplogroups. However, this issue can be resolved manually by examining particular variants within the CR.

The presence of both 499A and 16136C variants indicate an Asian B4b1 haplogroup while the presence of the 499A variant and the absence of the 16136C variant indicate a Native American B2 haplogroup. This demonstrates the capability of distinguishing Native American B2 haplogroups from Asian B4b1 haplogroups when only the CR is sequenced. High levels of variation and large numbers of private polymorphisms were observed in samples assigned to haplogroup B2. This indicates additional branches within B2 may exist that are not presently identified, expanding our knowledge of the known genetic variability within this haplogroup.

STR analyses of the Y-chromosome are used in anthropological genetics to identify paternally inherited haplogroups. These are generally identified by examining the length of each repeat. However, with NGS technologies, not only can the length of these alleles be learned but the sequence motif is revealed as well. These isoalleles and repeat pattern variants can be informative in population genetics as they have been suggested to be populationally distributed. The same sequence data can be learned with autosomal STRs as well, though these are not sex-specific. Nucleotide sequence data in conjunction with length variant data can dramatically increase our ability to genetically characterize populations, to reveal Y-chromosome haplogroup diversity, and opens a new door for the exploration of isoalleles in population genetics.

STR analyses revealed five autosomal STRs and four Y-STRs with an increase in the number of alleles when sequence-based alleles, produced from NGS, are considered in conjunction with length-based alleles, produced from traditional methods. The D12S391, D2S1338, vWA, D2S441, D7S820, DYS390, DYS438, DYS448, and DYS635 loci all

demonstrated sequence-based allele variance while the FGA, D16S539, and DYS392 loci did not demonstrate additional variation from sequence information.

Statistically significant differences among the frequency distributions of the length-based and sequence-based alleles of Native American samples and Asian samples were found for five loci: D2S441, D7S820, vWA, DYS392, and DYS635. Three of these loci, D2S441, D7S820, and DYS392 did not produce any further population differentiation when using SB alleles. Two of these loci, vWA and DYS635, produced higher significance values with the sequence-based alleles compared to the length-based alleles. This indicates these loci can be populationally informative in anthropological research as they have demonstrated statistical differences among groups that exceed the differences observed by length alleles alone. The frequency distributions of sequence-based alleles of these groups was significant at the D2S1338 locus, indicating sequence-based alleles can be populationally informative at this locus.

The research presented here demonstrates the increased genetic characterization of populations using next-generation sequencing as opposed to traditional methods. NGS technologies have progressed over the last decade to provide more affordable and accessible high-throughput options for all laboratories. Full mitogenome sequencing with NGS can produce more accurate mitochondrial haplogrouping and provide more informative haplotype information than traditional Sanger sequencing of small regions of the mitogenome. Autosomal STRs and Y-STRs sequenced via NGS can illuminate increased levels of genetic diversity not seen using traditional PCR-CE methods. The sequence-based allelic diversity has been shown to be populationally informative for several loci and useful in population genetics. These are some of the fundamental areas in which anthropological genetics can advance using next-generation sequencing technologies.

Works Cited

Achilli, A., Perego, U.A., Bravi, C.M., Coble, M.D., Kong, Q.P., Woodward, S.R.[…] (2008). The phylogeny of the four pan-American mtDNA haplogroups: Implications for evolutionary and disease studies. PLoS One 3, e1764.

Achilli, A., Perego, U.A., Lancioni, H., Oliveri, A., Gandini, F., Kashani, B.H., […] (2013). Reconciling migration models to the Americas with the variation of North American native mitogenomes. P Natl Sci USA, 110, 14308-14313.

Alvarez-Cubero, M.J., Saiz, M., Martínez-García, B., Sayalero, S.M., Entrala, C., Lorente, J.A., & Martinez-Gonzales, L.J. (2017). Next generation sequencing: an application in forensic sciences? Ann Human Biol, 44, 581-592.

Aly, S.M., & Sabri, D.M. (2015). Next generation sequencing (NGS): a golden tool in forensic toolkit. Arch Forensic Med Criminol, 65, 260-271.

Ambrosio, B., Dugoujon, J.M., Hernández, C., De La Fuente, D., Donzález-Martín, A., Fortes-Lima, C.A., Novelletto, J.N. […] (2010). The Andalusian population from Huelva reveals a high diversification of Y-DNA paternal lineages from haplogroup E: Identifying human male movements within the Mediterranean space. Ann Hum Biol, 37, 86-107.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J. […] (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457-465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Turnbull, R.N., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet, 23, 147.

Applied Biosystems. (2010). 7500/7500 Fast Real-Time PCR System Getting Started Guide for Standard Curve Experiments.

Applied Biosystems. (2011). SOLiD™ system accuracy with the exact call chemistry module. 5500 Series SOLiD System. Life Technologies. Carlsbad, CA. www.lifetechnologies.com

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y. […] (2010). Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet, 87, 341-353.

Ballantyne, K.N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S.B., Ralf, A., Vermeulen, M., de Knijff, & Kayser, M. (2012). A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Sci Int Genet, 6, 208-218.

Beckman Coulter. (2009). Xtra Performance Post-PCR Cleanup: Agencourt AMPure XP System PCR Purification System. DS-12722. www.beckmancoulter.com.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, H.P. […] (2008.) Accurate whole genome sequencing using reversible terminator chemistry. Nature, 456, 53-59.

Berglund, E.C., Kiialainen, A., & Syvänen, A.C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. Investig Genet, 2, 23.

Bornman, D.M., Hester, M.E., Schuetter, J.M., Kasoji, M.D., Minard-Smith, A., Barden, C.A., Nelson, S.C. […] (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotec Rapid Dispatchces, 1-6.

Børsting, C. & Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. Forensic Sci Int-Gen, 18, 78-89.

Budowle, B., Masibay, A., Anderson, S.J., Barna, C., Biega, L., Brenneke, S., Brown, B.L. […] (2001). STR primer concordance study. Forensic Sci Int, 124, 47-54.

Budowle, B., Adamowicz, M., Aranda, X.G., Barna, C., Chakraborty, R., Cheswick, D., Dafoe, B. […] (2005). Twelve short tandem repeat loci and Y chromosome haplotypes: Genetic analysis on populations residing in North America. Forensic Sci Int, 150, 1-15.

Butler, J.M. (2007). Short tandem repeat typing technologies used in human identity testing. BioTechniques: Suppl, 43, ii-v.

Caratti, S., Turrina, S., Ferrian, M., Cosentino, E., & De Leo, D. (2015). MiSeq FGx sequencing system: A new platform for forensic genetics. Forensic Sci Intl: Genet Suppl Series, 5, e98-e100.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A. […] (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci, 106, 19096-19101.

Coble, M.D., Loreille, O.M., Wadhams, M.J., Edson, S.M., Maynard, K., Meyer, C.E., Niederstätter […] (2009). Mystery solved: The identification of the two missing Romanov children using DNA analysis. PLoS One, 4, e4838.

Coia, V., Cipollini, G., Anagnostou, P., Maixner, F., Battaggia, C., Brisighelli, F., Gomez-Carballa, A. […] (2016). Whole mitochondrial DNA sequencing in Alpine populations and the genetic history of the Neolithic Tyrolean Iceman. Sci Rep-UK, 6, 18932.

Crawford, M.H. (1998). The origins of Native Americans: Evidence from anthropological genetics. Cambridge: Cambridge University Press.

Crawford, M.H., Rubicz, R.C., & Zlojutro, M. (2010). Origins of Aleuts and the genetic structure of populations of the archipelago: Molecular and archaeological perspectives. Hum Biol, 82, 695-717.

Crawford, M.H., & Beaty, K.G. (2013). DNA fingerprinting in anthropological genetics: past, present, future. Invest Genet, 4, 23.

Cruz, C., Ribeiro, T., Vieira-Silva, C., Lucas, I., Espinheira, R. & Geada, H. (2004). vWA STR locus structure and varaiability. Prog Forensic Genet, 10, 248-250.

Cui, Y., Lindo, J., Hughes, C.E., Johnson, J.W., Hernandez, A.G., Kemp, B.M., Ma, J. […] (2013). Ancient DNA analysis of mid-Holocene individuals from the northwest coast of North America reveals different evolutionary oaths for mitogenomes. PLoS One, 8, e66948.

Dauber, E-M., Bar, W., Klintschar, M., Neuhuber, F., Parson, W., Mueller-van der Spruit, E., & Mayr, W.R. (2004). New sequence data of allelic variants at the STR loci ACTBP2 (SE33), D21S11, FGA, vWA, CSF1PO, D2S1338, D16S539, D18S51, and D19S433 in Caucasoids. Prog For Genet, 1261, 191-193.

Dauber, E-M., Kratzer, A., Neuhuber, F., Parson, W., Klintschar, M., Bär, W., & Mayr, W.R. (2012). Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions. Forensic Sci Int Genet, 6, 381-386.

Davies, M.J., Smethurst, D.E., Howard, K.M., Todd, M., Higgins, L.M., & Bruce, I.J. (1997). Improved manufacture and application of an agarose magnetizable solid-phase support. Appl Biochem Biotech, 68, 95-112.

Derenko, M., Grzybowski, T., Malyarchuk, B.A., Czarny, J., Miścicka-Sliwka, D., & Zakharov, I.A. (2001). The presence of mitochondrial haplogroup X in Altaians from South Siberia. Am J Hum Genet, 69, 237-241.

Derenko, M., Grzybowski, T., Malyarchuk,A., Dambueva, I.K., Denisova, G.A., Czarny, J., Dorzhu, C.M. […] (2003). Diversity of mitochondrial DNA lineages in South Siberia. Ann Hum Genet, 67, 391-411.

Derenko, M., Malyarchuk, B., Denisova, G., Perkova, M., Rogalla, U., Grzybowski, T., […] (2012). Complete mitochondrial DNA analysis of eastern Eurasian haplogroups rarely found in populations of northern Asia and eastern Europe. PLoS One 7, e32179.

Dos Santos, S.E.B., Ribeiro-Rodrigues, E.M., Ribeiro-Dos-Santos, A.K.C., Hutz, M.H., Tovo-Rodrigues, L., Salzano, F.M., & Callegari-Jacques, S.M. (2009). Autosomal STR analyses in Native Amazonian tribes suggest a population structure driven by isolation by distance. Hum Biol, 81, 71-88.

Elkin, C., Kapur, H., Smith, T., Humphries, D., Pollard, M., Hammon, N., & Hawkins, T. (2001). Magnetic bead purification of labeled DNA fragments for high-throughput capillary

electrophoresis sequencing. Lawrence Berkeley National Laboratory. http://escholarship.org/uc/item/95x6q4pd

Ermini, L., Olivieri, C., Rizzi, E., Corti, G., Bonnal, R., Soares, P., Luciani, S. […] (2008). Complete mitochondrial genome sequence of the Tyrolean Iceman. Curr Biol, 18, 1687-1693.

Excoffier, L., & Lischer, H.E. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molec Ecol Res, 10, 564-567.

Ezewudo, M. & Zwick, M.E. (2013). Evaluating rare variants in complex disorders using next-generation sequencing. Curr Psychiat Rep 15, 349.

Fox, E.J., Reid-Bayliss, K.S., Emond, M.J., & Loeb, L.A. (2014). Accuracy of next generation sequencing platforms. Next Gener Seq Appl, 1, 1000106.

Garrido-Cardenas, J.A., Garcia-Maroto, F., Alvarez-Bermejo, J.A., & Manzano-Agugilaro, F. (2017). DNA sequencing sensors: An overview. Sensors (Basel), 3, 588. doi:10.3390/s17030588.

Gelardi, C., Rockenbauer, E., Dalsgaard, S., Børsting, C., Morling, N. (2014). Second generation sequencing of three STRs D3S1358, D12S391, and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci Int Genet, 12, 38-41.

Gettings, K.B., Aponte, R.A., Vallone, P.A., & Butler, J.M. (2015). STR allele sequence variation: current knowledge and future issues. Forensic Sci Int Genet, 19, 118-130.

Gettings, K.B., Kiesler, K.M., Faith, S.A., Montano, E., Baker, C.H., Young, B.A., Guerrieri, R.A., […] (2016). Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Sci Int Genet, 21, 15-21.

Goebel, T., Waters, M.R., & O'Rourke, D. (2008). The late Pleistocene dispersal of modern humans in the Americas. Science, 319, 1497-1502.

Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A., & Tishkoff, S.A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol, 24, 757-768.

Goodwin, S., McPherson, J.D., & McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet, 17, 333-351.

Green, E.D. (2001). Strategies for the systematic sequencing of complex genomes. Nat Rev Genet, 2, 573-583.

Grubwieser, P., Muhlmann, R., Niederstatter, H., Pavlic, M., & Parson, W. (2005). Unusual variant alleles in commonly used short tandem repeat loci. Int J Legal Med, 119, 164-166.

Guchelaar, H.-J., Gelderblom, H., van der Straaten, T., Schellens, J.H.M., & Swen, J.J. (2014). Pharmacogenetics in the cancer clinic: From candidate gene studies to next-generation sequencing. Clin Pharmacol Ther, 95, 383-385.

Guo, F. (2017). Population genetic data for 12 X-STR loci in the Northern Han Chinese and StatsX package as tools for population statistics on X-STR. Forensic Sci Int Genet, 26, e1-e8.

Guo, S.W., & Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, 48, 361-372.

Gužvić, M. (2013). The history of DNA sequencing. J Med Biochem. 32, 301-312.

Hadidi, A., & Candresse, T. (2003). Polymerase Chain Reaction. Hadidi, A., Flores, R., Randles, J. & Semancik, (Eds.). Viroids: Properties, Detection, Diseases and their control. Callingwood, Australia: Csiro Publishing.

Harakalova, M., Mokry, M., Hrdlickova, B., Renkens, I., Duran, K., van Roekel, H., Lansu, N. […] (2011.) Multiplexed array-based and in-solution genomic enrichment for flexible and cost-effective targeted next-generation sequencing. Nat Protoc, 6, 1870-1886.

He, G., Li, Y., Wang, Z., Liang, W., Luo, H., Liao, M., Zhang, J […] (2017). Genetic diversity of 21 autosomal STR loci in the Han population from Sichuan province, Southwest China. Forensic Sci Intl Genet, 31, e33-e35.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. Biotechniques, 56, 61-77.

Heather, J.M, & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107, 1-8.

Holland, M.M., McQuillan, M.R., & O'Hanlon, K.A. (2011). Second generation sequencing allows for mtDNA mixture and deconvolution and high resolution detection of heteroplasmy. Croat Med J, 52, 299-313.

Holt, R.A., & Jones, S.T.M. (2008). The new paradigm of flow cell sequencing. Genome Res, 18, 839-846.

Huang, Y.F., Chen, S.C., Chiang, Y.S., Chen, T.H., & Chiu, K.P. (2012). Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol, 6 (Suppl 2), S10.

Huber, N., & Parson, W. SAM 2- A second generation mitochondrial DNA database search algorithm for unbiased sequence queries and alignment harmonization. In prep.

Illumina. (2015). ForenSeq™ DNA Signature Prep, Reference Guide, v01.

Illumina. (2016). ForenSeq™ Universal Analysis Software Guide, v01.

Illumina. (2017). Prepare Library. Best Practices for Standard and Bead-Based Normalization in Nextera XT DNA Library Preparation Kits.

Illumina, MiSeq System. (2016). Denature and Dilute Libraries Guide, v2.

Iozzi, S., Carboni, I., Contini, E., Pescucci, C., Frusconi, S., Nutini, A.L.., Torricelli, F. […] 2015. Forensic genetics in NGS era: New frontiers for massively parallel typing. Forensic Sci Int Genet Suppl, Series 5, e418-e419.

Irwin, J., Saunier, J.L., Strouss, K.M., Sturk, K.A., Diegoli, T.M., & Just, R.S. (2007). Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. Forensic Sci Int Genet, 1, 154-157.

Jeffreys, A.J., Wilson, V., & Thein, S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. Nature, 314, 67-73.

Just, R.S., Irwin, J.A., & Parson, W. (2015). Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int Genet, 18, 131-139.

Just, R.S., Moreno, L.I., Smerick, J.B., & Irwin, J.A. (2017). Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. Forensic Sci Int Genet, 28, 1-9.

Kashani, B.H., Perego, U.A., Olivieri, A., Angerhofer, N., Gandini, F., Carossa, V., Lancioni, H. […] (2012). Mitochondrial haplogroup C4c: A rare lineage entering American through the ice-free corridor? Am J Phys Anthropol, 147, 35-39.

Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics, and molecular biology. Nat Rev Genet, 12, 179-192.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., & Vogelstein, B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci, 108, 9530-9535.

King, J.L., LaRue, B.L., Novroski, N., Stoljarova, M., Bum Seo, S., Zeng, X., Warshauer, D.H. […] (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet, 12, 128-135.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing-concepts and limitations. Bioessays, 32, 524-536.

Kline, M.C., Hill, C.R., Decker, A.E. & Butler, J.M. (2011). STR sequence analysis for characterizing normal: variant, and null alleles. Forensic Sci Int Genet, 5, 329-332.

Kolman, C.J., Sambuughin, N., Bermingham, E. (1996). Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics, 142, 1321-1334.

Kulski, J. K. (2016). Chapter 1: Next-generation sequencing- An overview of the history, tools, and "omic" applications. Kulski, J, K., (Ed.) Next-generation sequencing- Advances, applications and challenges. Intech Open.

Kumar, S., Bellis, C., Zlojutro, M., Melton, P.E., Blangero, J., & Curran, J.E. (2011). Large scale mitochondrial sequencing in Mexican Americans suggest a reappraisal of Native American origins. BMC Evol Biol, 11, 293.

Kwon, S.Y., Lee, H.Y., Kim, E.H., Lee, E.Y., & Shin, K.-J. (2016). Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing. Forensic Sci Int Genet, 25, 132-141.

Kwong, J.C., McCallum, N., Sintchenko, V., & Howden, B.P. (2015). Whole genome sequencing in clinical and public health microbiology. Pathol, 47, 199-210.

Lam, H.Y., Clark, M.J., Chen, Ru., Chen Ro., Natsoulis, G., O'Huallachain, M., Dewey, F.E. […]. (2012). Performance comparison of whole-genome sequencing platforms. Nat Biotechnol, 30, 78-82.

Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J. […] (2003). A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis, 24, 3769-3777.

Levison, P.R., Badger, S.E., Dennis, J., Hathi, P., Davies, M.J., Bruce, I.J., & Schimkat, D. (1998). Recent developments of magnetic beads for use in nucleic acid purification. *J* Chromatogr A, 816, 107-111.

Li, H., & Homer, N. (2011). A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform, 11, 473-483.

Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A. […] (2014) Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol, 32, 915-925.

Liang, F., Liu, C., & Carroll, R. (2010). Advanced Markov Chain Monte Carlo Methods: Learning from past samples. West Sussex: John Wiley & Sons, Inc.

Lim, S.K., Xue, Y., Parkin, E.J., & Tyler-Smith, C. (2007). Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals. Intl J Legal Med, 121, 124-127.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D. […]. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol, 2012 doi: 10.1155/2012/251364.

Llamas, B., Fehren-Schmitz, L., Valverde, G., Soubrier, J., Mallick, S., Rohland, N., Nordenfelt, S. […] (2016). Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv, 2, e15011385.

Lopopolo, M., Børsting, C., Pereira, V., Morling, N. (2016). A study of the peopling of Greenland using next-generation sequencing of complete mitochondrial genomes. Am J Phys Anthropol, 161, 689-704.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Ann Rev Genomics Hum Genet, 9, 387-402.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature, 470, 198-203.

Mardis, E.R. (2013). Next-generation sequencing platforms. Ann Rev Anal Chem, 6, 287-303.

Marks, S. J., Levy, H., Martinez-Cadenas, C., Montinaro, F., & Capelli, C. (2012). Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. Mol Ecol, 21, 4958-4969.

Maruki, T., & Lynch, M. (2015). Genotype-frequency estimation from high-throughput sequencing data. Genetics, 201, 473-486.

Matisoo-Smith, E., Gosling, A.L., Platt, D., Kardailsky, O., Prost, S., Cameron-Christie, S., Collins, C.J. […] (2018). Ancient mitogenomes of Phoenicians from Sardinia and Lebanon: A story of settlements, integration, and female mobility. PLos One, 13, e0190169.

Matullo, G., Di Gaetano, C., & Guarrera, S. (2013). Next generation sequencing and rare genetic variants: From human population studies to medical genetics. Environ Mol Mutagen, 54, 518-532.

Merriwether, D.A., Hall, W.W., Vahlne, A., & Ferrell, R.E. (1996). mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. Am J Hum Genet, 59, 204-212.

Metzker, M.L. (2010). Sequencing technologies- the next generation. Nat Rev Genet, 11, 31-46.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A., Hosseini, S., Brandon, M. […] (2003). Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci, 100, 171-176.

Mitchell, R.J., Reddy, B.M., Campo, D., Infantino, T., Kaps, M., Crawford, M.H. (2006). Genetic diversity within a caste population of India as measured by Y-chromosome haplogroups and haplotypes: Subcastes of the Golla of Andhra Pradesh. Am J Phys Anthropol, 130, 385-393.

Mitra, R.D., Shendure, J., Olejnik, J., Olejnik, E.K., & Church, G.M. (2003). Fluorescent in situ sequencing on polymerase colonies. Anal Biochem, 320, 55-65.

Morozova, O., & Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics, 92, 255-264.

Nagle, N., Ballantyne, K.N., van Oven, M., Tyler-Smith, C., Zue, Y., Taylor, D., Wilcox, S. […] (2015). Antiquity and diversity of aboriginal Australian Y-chromosomes. Am J Phys Anthropol, 159, 367-381.

Nagle, N., van Oven, M., Wilcox, S., van Holst Pellekaan, S., Tyler-Smith, C., Xue, Y., Ballantyne, K.N. […] (2017). Aboriginal Australian mitochondrial genome variation- an increased understanding of population antiquity and diversity. Sci Rep-UK, 7, 43041.

Nargessi, R.D. (2005). Magnetic isolation and purification of nucleic acids. United States patent US 6855499 B1. Cortex Biochem, Inc.

Neparáczki, E., Kocsy, K., Tóth, G.E., Maróti, Z., Kalmár, T., Bihari, P., Nagy, I. […] (2017). Revising mtDNA haplotypes of the ancient Hungarian conquerors with next generation sequencing. PLoS One, 12, e0174886.

Neverov, A., Chumakov, K., & Purcell, R.H. (2010). Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. Proc Natl Acad Sci, 107, 20063-20068.

New England Biosystems, Inc. (2016.) NEBNext® for Illumina: NGS sample preparation. Version 4.0. www.neb.com.

Novroski, N.M.M, King, J.L., Churchill, J.D., Seah, L.H., & Budowle, B. (2016). Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci Int Genet, 25, 214-226.

Nuñez, C., Baeta, M., Sosa, C., Casalod, Y., Ge, J., Budowle, B., Martínez-Jarreta, B. (2010). Reconstructing the population history of Nicaragua by means of mtdna, Y-chromosome STRs, and autosomal STR markers. Am J Phys Anthropol, 143, 591-600.

Nuwaysir, E.F., Huang, W., Albert, T.J., Singh, J., Nuwaysir, K., Pitas, A., & Richmond, T. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res, 12, 1749-1755.

O'Rourke, D.H., & Raff, J.A. (2010). The human genetic history of the Americas: The final frontier. Curr Biol 20, R202-R207.

Olofsson, J.K., Pereira, V., Børsting, C., & Morling, N. (2015). Peopling of the North circumpolar region- Insights from Y chromosome STR and SNP typing of Greenlanders. PLoS One, 10, e0116573.

Park, J.Y., Kricka, L.J., & Fortina, P. (2013). Next-generation sequencing in the clinic. Nat Biotechnol, 31, 990-992.

Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S.M., Souto, L., Fendt, L. […] (2013). Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet, 7, 543-549.

Parson, W., Ballard, D., Budowle, B., Butler, J.M., Gettings, K.B., Gill, P., Gusmão, L […] (2016). Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet, 22, 54-63.

Peck, M.A., Sturk-Andreaggi, K., Thomas, J.T., Oliver, R.S., Barritt-Ross, S., & Marshall, C. (2018). Developmental validation of a Nextera XT mitogenome Illumina MiSeq sequencing method for high-quality samples, Forensic Sci Int Genet, doi: https://doi.org/10.1016/j.fsigen.2018.01.004.

Phillips-Krawczak, C., Devor, E., Zlojutro, M., Moffat-Wilson, K., & Crawford, M.H. (2006). mtDNA variation in the Altai-Kizhi population of southern Siberia: A synthesis of genetic variation. Hum Biol, 78, 477-494.

Planz, J.V., Sannes-Lowery, K.A., Duncan, D.D., Manalili, S., Budowle, B., Chakraborty, R., Hofstadler, S.A., […] (2012). Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. Forensic Sci Int Genet, 6, 594-606.

Raff, J.A., & Bolnick, D. A. (2015). Does mitochondrial haplogroup X indicate ancient trans-Atlantic migration to the Americas? A critical re-evaluation. PaleoAmerica, 1, 297-304.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H.-V., Parik, J. […] (2003). Origin and diffusion of mtDNA haplogroup X. Am J Hum Genet, 73, 1178-1190.

Ring, J.D., Sturk-Andreaggi, K., Peck, M.A., & Marshall, C. (2017). A performance evaluation of Nextera XT and KAPA HyperPlus for rapid Illumina library preparation of long-range mitogenome amplicons. Forensic Sci Intl Genet 29, 174-180.

Rizzo, J.M., & Buck, M.J. (2012). Key principles and clinical applications of "Next-generation" DNA sequencing. Cancer Prev Res, 5, 887-900.

Röck, A.W., Dür, A., van Oven, M., & Parson, M. (2013). Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). Forensic Sci Int Genet, 7, 601-609.

Rockenbauer, E., Hansen, S., Mikkelsen, M., Børsting, C., & Morling, N. (2014). Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. Forensic Sci Int Genet, 8, 68-72.

Rousset, F. (2008). GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Molec Ecol Res, 8, 103-106.

Rubicz, R.C., Melton, P.E., & Crawford, M.H. (2006). Molecular markers in anthropological genetic studies. Crawford, M.H., (ed.) Anthropological genetics: Theory, methods, and applications. Cambridge: Cambridge University Press.

Rubicz, R., Melton, P., Spitsyn, V., Sun, G., Deka, R., Crawford, M.H. (2010). Genetic structure of native circumpolar populations based on autosomal, mitochondrial, and Y chromosome DNA markers. Am J Phys Anthropol, 143, 62-74.

Rubicz, R., Zlojutro, M., Sun, G., Spitpsyn V., Deka, R., Young, K.L., & Crawford, M.H. (2010). Genetic architecture of a small, recently aggregated Aleut population: Bering Island, Russia. Hum Biol, 82, 719-736.

Ruitberg, C.M., Reeder, D.J., & Butler, J.M. (2001). STRBase: A short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res, 29, 320-322.

Saiyed, Z.M., Bochiwal, C., Gorasia, H., Telang, S.D., & Ramchand, C.N. (2006). Application of magnetic particles (Fe_3O_4) for isolation of genomic DNA from mammalian cells. Anal Biochem, 356, 306-308.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 74, 5463-5467.

Schanfield, M. (2007). Applications of molecular genetics to forensic sciences. In: Crawford M (Ed.) Anthropological genetics: Theory, methods, and applications (p. 235-276). New York: Cambridge University Press.

Scheible, M., Loreille, O., Just, R., & Irwin, J. (2014). Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci Int Genet, 12, 107-119.

Sharma, V., Chow, H.Y., Siegel, D., & Wurmbach, E. (2017). Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx™. PLoS One, 12, e0187932.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol, 26, 1135-1145. doi: 10.1038/nbt1486.

Shendure, J.A., Porreca, G.J., & Church, G.M. (2008). Overview of DNA sequencing strategies. Curr Protoc Mol Biol, 81, 7.1.1-7.1.11.

Shields, G.F., Hecker, K., Voevoda, M.I., & Reed, J.K. (1992). Absence of the Asian-specific region V mitochondrial marker in native Beringians. Am J Hum Genet, 50, 758-765.

Shrivastava, P., Jain, T., & Trivdei, V. B. (2017). Structure and genetic relationship of five populations from central India based on 15 autosomal STR loci. Ann Hum Biol, 44, 74-86.

Soares, P. Ermini, L., Thomson, N., Mormina, M. Rito, T., Rohl, A., Salas, A. […] (2009). Correcting for purifying selection: An improved human mitochondrial molecular clock. Am J Hum Genet, 84, 740-759.

Sokal, R.R., & Rohlf, F.J. (2012). Biometry: The principles and practice of statistics in biological research, 4th ed. New York: W.H. Freeman Company.

Sosa, M.X., Ashok Sivakumar, I.K., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K.M. […] (2012). Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol, 8, e1002737.

Starikovskaya, E.B., Sukernik, R.I., Schurr, T.G,m Kogelnik, A.M., & Wallace, D.C. (1998). mtDNA diversity in Chukchi and Siberian Eskimos: Implications for the genetic history of Ancient Beringia and the peopling of the New World. Am J Hum Genet, 63, 1473-1491.

Starikovskaya, E.B., Sukernik, R.I., Derbeneva, O.A., Volodko, N.V., Ruiz-Pesini, E., Torroni, A. […] (2005). Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. Ann Human Genet, 69, 67-89.

Stewart, J.B., & Chinnery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. Nat Rev Genet, 16, 530-542.

Sturk-Andreaggi, K., Peck, M.A., Boysen, C., Dekker, P., McMahon, T.P., & Marshall, C.K. (2017). AQME: A forensic mitochondrial DNA analysis tool for next-generation sequencing data. Forensic Sci Int Genet, 21, 189-197.

Sukernik, R.I., Schurr, T.G., Stariovskaya, Y.B., & Wallace, D.C. (1996). Mitochondrial DNA variation in native Siberians, with special reference to the evolutionary history of American Indians: Studies on restriction endonuclease polymorphism. Genetika, 32, 432-439.

Suzuki, M. & Greally, J.M. (2013). Genome-wide DNA methylation analysis using massively parallel sequencing technologies. Semin Hematol, 50, 70-77.

Tackney, J.C., Potter, B.A., Raff, J., Powers, M., Watkins, W.S., Warner, D., Reuther, J.D. […] (2015). Two contemporaneous mitogenomes from terminal Pleistocene burials in eastern Beringia. Proc Natl Acad Sci, 112, 13822-13838.

Tan, S.C., & Yiap, B.C. (2009). DNA, RNA, and Protein Extraction: The past and the present. J Biomed and Biotechnol, 2009, doi:10.1155/2009/574398.

Tanaka, M., Hayakawa, M., & Ozawa, T. (1996). Automated sequencing of mitochondrial DNA, Method Enzymol 264, 407-421.

Tanaka, M., Cabrera, V.M., González, A.M., Larruga, J.M., Takeyasu, T., Fuku, N. […] (2004). Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res, 14, 1832-1850.

Torroni, A., Schurr, T.G., Cabell, M.F., Brown, M.D., Neel, J.V., & Larsen, M. (1993a). Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet, 53, 563-590.

Torroni, A., Sukernik, R.I., Schurr, T.G., Starikorskaya, Y.B., Cabell, M.F., Crawford, M.H. […]. (1993b) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet 53, 591-608.

Torroni, A., Chen, Y.-S., Semino, O., Santachiara-Beneceretti, A.S., Scott, C.R., Lott, M.T., Winter, M. […] (1994). mtDNA and Y-chromosome polymorphisms in four Native American populations from Southern Mexico. Am J Hum Genet, 54, 303-318.

Tucker, T., Marra, M., & Friedman, J.M. (2009). Massively parallel sequencing: The next big thing in genetic medicine. Am J Hum Genet, 85, 142-154.

Tzvetkov, M. & von Ahsen, N. (2012). Pharmacogenetic screening for drug therapy: From single gene markers to decision making in the next generation sequencing era. Pathology, 44, 166-180.

van der Gaag, K.J., de Leeuw, R.H., Hoogenboom, J., Patel, J., Storts, D.R., Laros, J.F.J., & de Knijff, P. (2016). Massively parallel sequencing of short tandem repeats- Population data and mixture analysis results for the PowerSeq system. Forensic Sci Int Genet, 24, 86-96.

van Oven, M. (2015). PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Sci Int Genet 5, e392-e394.

Veeramah, K.R. & Hammer, M.F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet, 15, 149-162.

Wallace, D.C, Garrison, K., & Knowler, W.C. (1985). Dramatic founder effects in Amerindian mitochondrial DNAs. Am J Phys Anthropol, 68, 149-155.

Wang, M., Wang, Z., Zhang, Y., He, G., Liu, J., Hou, Y. (2017). Forensic characteristics and phylogenetic analysis of two Han populations from the southern coastal regions of China using 27 Y-STR loci. Forensic Sci Int Genet, 31, e17-e23.

Warshauer, D.H., Churchill, J.D., Novroski, N., King, J.L., Budowle, B. (2015). Novel Y-chromosome short tandem repeat variants detected through the use of massively parallel sequencing. Genomics Proteomics Bioinformatics, 13, 250-257.

Wendt, F.R., Churchill, J.D., Novroski, N.M., King, J.L., Ng, J., Oldt, R.F., McCulloh, K.L. […] (2016). Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system. Forensic Sci Int Genet, 24, 18-23.

Xue, Y., & Tyler-Smith, C. (2010). The hare and the tortoise: one small step for four SNPs, one giant leap for SNP-kind. Forensic Sci Int Genet, 4, 59-61.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C. […] (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol, 15, 1453-1457.

Yang, Y., Xie, B., & Yan, J. (2014). Application of next-generation sequencing technology in forensic science. Genomics Proteomics Bioinformatics, 12, 190-197.

Young, K.L., Sun, G., Deka, R., & Crawford, M.H. (2011). Autosomal short tandem repeat genetic variation of the Basques in Spain. Croat Med J, 52, 372-383.

Yuan, A., & Bonney, G. (2003). Exact test of Hardy-Weinberg equilibrium by Markov chain Monte Carlo. Math Med Biol, 20, 327-340.

Zeggini, E. (2011). Next-generation association studies for complex traits. Nat Genet, 43, 287-288.

Zeng, X., King, J.L., Stoljarova, M., Warshauer, D.H., LaRue, B.L., Sajantila, A., Patel, J. […] (2015). High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet, 16, 38-47.

Zhao, X., Ma, K., Li, H., Cao, Y., Liu, W., Zhou, H., & Ping, Y. (2015). Multiplex Y-STRs analysis using the ion torrent personal genome machine (PGM). Forensic Sci Int Genet, 19, 192-196.

Zhao, X., Hui, L., Wang, Zz., Cao, Y., & Liu, W. (2016). Massively parallel sequencing of 10 autosomal STRs in Chinese using the ion torrent personal genome machine (PGM). Forensic Sci Int Genet, 25, 34-38.

Zlojutro, M., Roy, R., Palikij, J., & Crawford, M.H. (2006). Autosomal STR variation in a Basque population: Vizcaya Province. Hum Biol, 78, 599-618.

Appendix A: Samples with Contributors and Extraction Method

Sample	Self-Identified Population / Metadata	Contributor	Sample Type	AFDIL DNA Extraction Method	Extraction Instrument
CHN007	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
CHN031	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
CHN094	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
CHN120	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
CHN129	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
CHN157	Chinese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
ILH012	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH017	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH070	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH071	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH074	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH084	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH087	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
ILH097	Illinois Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
JPN050	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN063	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN080	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN138	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN199	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN200	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN207	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN260	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN274	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
JPN275	Japanese	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
NYAS062	New York Asian American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
NYAS078	New York Asian American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
OHHis035	Ohio Hispanic	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
OHHis068	Ohio Hispanic	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
OHHis103	Ohio Hispanic	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
OHHis116	Ohio Hispanic	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL005	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)

PHL012	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL035	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL050	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL052	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL055	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL061	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL071	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL079	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL084	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL088	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL097	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL098	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL100	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL106	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL109	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL110	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL140	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL142	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL145	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
PHL154	Filipino	DNA Diagnostics Center, Fairfield, Ohio	Buccal swab	DNA IQ System	Biomek 2000 (Beckman Coulter)
SDHis001	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDHis020	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA003	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA029	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA035	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA052	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA055	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA058	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA060	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	

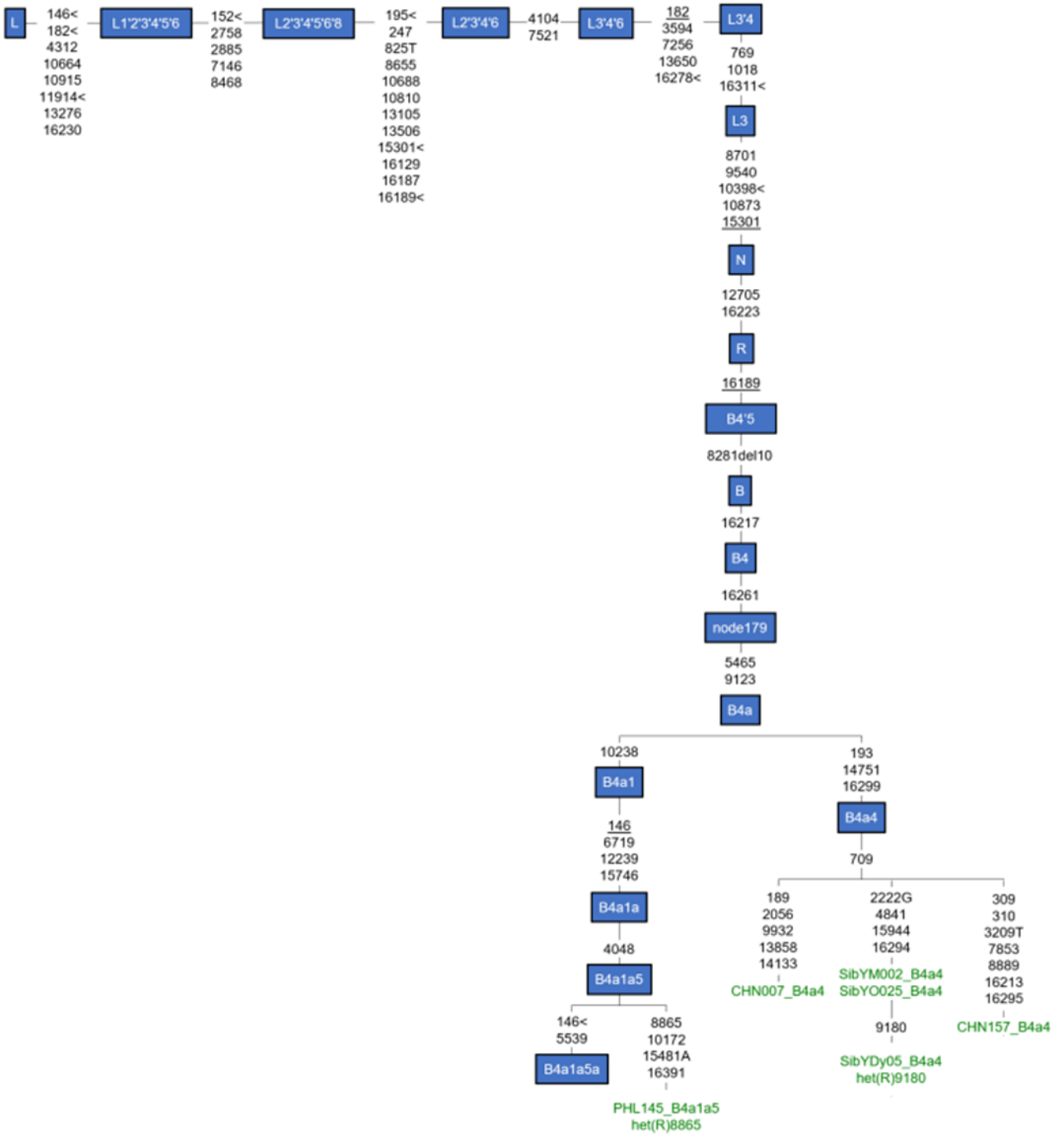
SDNA088	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA106	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA126	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA127	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA129	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA130	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SDNA150	South Dakota Native American	South Dakota Department of Public Safety	Bloodstain	Received at AFDIL as extracts	
SibA009	Siberian - Altai Mountains	Michael Crawford	DNA extract	Received at AFDIL as extracts	
SibA096	Siberian - Altai Mountains	Michael Crawford	DNA extract	Received at AFDIL as extracts	
SibYDe54	Siberian Yakut - Debdirge	Michael Crawford / Larissa Nichols	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
SibYDy05	Siberian Yakut - Dygdal	Michael Crawford / Larissa Nichols	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
SibYM002	Siberian Yakut - Mukuchu Kobiyskiy region	Michael Crawford / Larissa Nichols	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
SibYO025	Siberian Yakut - Orto-Surt Gorny region	Michael Crawford / Larissa Nichols	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
TXHis033	Texas Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	STARlet (Hamilton)
TXHis117	Texas Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	STARlet (Hamilton)
TXHis135	Texas Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	STARlet (Hamilton)
TXHis167	Texas Hispanic	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	STARlet (Hamilton)
VTAS001	Vermont Asian American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
VTAS016	Vermont Asian American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA007	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA037	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA050	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA062	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA065	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)
WANA093	Washington Native American	Department of Defense Serum Repository	Blood serum	QIAamp 96 DNA Blood Kit	Bio-Robot 9604 (Qiagen)

Appendix B: mtDNA Sequence Metrics for Each Sample

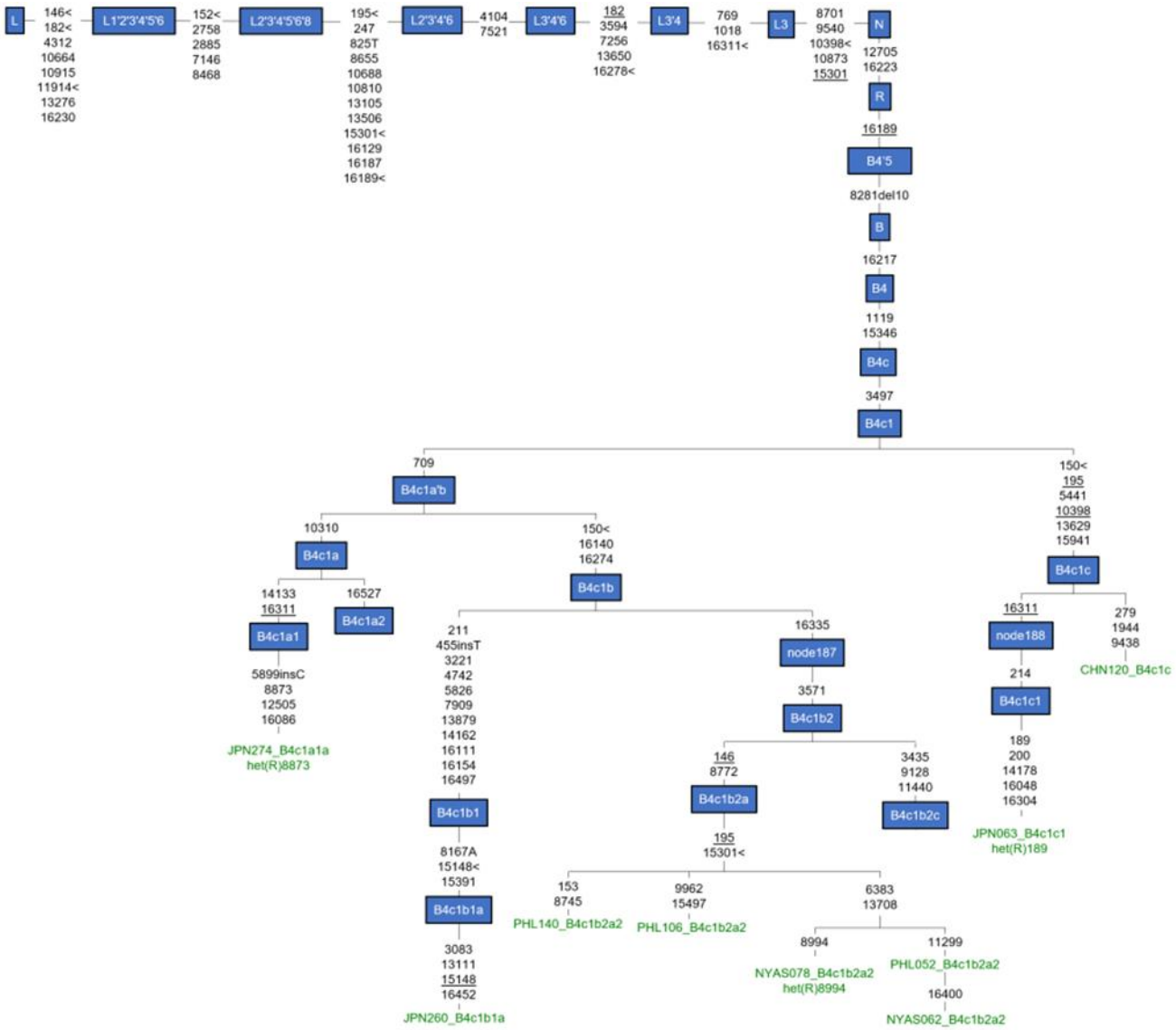
Sample	# Reads	# Reads After Trimming	# Reads Mapped to rCRS	# Bases Reported	Avg Coverage	Avg For/Rev Balance	Avg For/Rev Balance (excluding 309, 16182, 16189 length variants)	Avg VF (excluding Het, Indels, & filtered variants)
CHN007	148652	109362	103894	16569	581.1	0.4284	0.4717	98.6288
CHN094	120845	92103	85320	16569	475.3	0.4207	0.4720	99.6163
CHN120	96508	70901	65047	16569	362	0.4102	0.4638	99.7501
CHN129	92190	65165	57346	16569	325.1	0.4509	0.4622	99.6692
CHN157	97380	63763	56714	16569	321.7	0.4373	0.4466	99.4801
ILH012	118917	91714	86071	16569	486.2	0.4466	0.4719	99.6718
ILH017	107360	81377	75841	16569	428.5	0.4113	0.4762	99.6625
ILH070	126254	93577	87054	16569	493.3	0.4268	0.4599	99.6896
ILH071	124401	93293	85976	16569	487.5	0.4281	0.4673	99.6624
ILH074	145462	94067	83814	16569	478	0.4213	0.4735	99.7526
ILH084	135196	104364	96196	16569	543	0.4587	0.4657	99.7884
ILH087	118417	81812	71203	16569	402.7	0.4081	0.4633	99.7936
ILH097	122438	95497	91291	16569	512.8	0.4134	0.4743	99.6667
JPN063	130768	91597	83343	16569	454.3	0.4193	0.4765	99.6951
JPN080	105491	75262	69607	16569	382.8	0.3830	0.4624	99.4476
JPN138	131333	96340	86912	16569	489.2	0.4149	0.4797	99.5522
JPN260	117136	87943	77935	16569	447	0.4200	0.4653	99.5907
JPN274	115524	83349	78153	16569	436.3	0.3863	0.4657	99.3845
JPN275	123914	87125	79121	16569	444.8	0.4083	0.4718	99.6553
NYAS062	141228	104371	94585	16569	522.3	0.3998	0.4621	99.6991
NYAS078	132157	94179	86558	16569	481.6	0.4043	0.4625	99.6842
OHHis035	92679	70394	65636	16569	371.7	0.4452	0.4610	99.6132
OHHis068	118842	89976	84174	16569	474.9	0.3788	0.4542	99.7030
OHHis103	92803	71740	67477	16569	379.5	0.3623	0.4591	99.3709
OHHis116	125553	95774	89183	16569	501.5	0.3865	0.4673	99.7147
PHL012	106251	83890	78609	16569	442.1	0.4346	0.4674	99.5281
PHL052	124313	94228	86577	16569	479.2	0.3832	0.4601	99.5991
PHL106	121339	89777	82802	16569	465.3	0.4254	0.4681	99.6219
PHL109	106997	81710	76409	16569	430.7	0.3716	0.4558	99.6566
PHL110	115695	83666	77973	16569	435.1	0.4381	0.4660	99.6560
PHL140	131547	96870	89826	16569	492.6	0.4271	0.4647	99.5964
PHL142	135224	99147	92687	16569	517.6	0.3780	0.4362	99.3322
PHL145	144784	107198	100367	16569	556.1	0.4466	0.4698	99.6879
PHL154	131671	101793	94966	16569	535.4	0.4153	0.4627	99.6194
SDNA029	115575	87703	82533	16569	465.8	0.4021	0.4761	99.7188
SDNA035	116104	85875	80918	16569	460.3	0.4205	0.4675	99.6228
SDNA060	106720	75105	69847	16569	384.3	0.4239	0.4627	99.6549

SDNA126	119435	91101	83833	16569	473.9	0.4202	0.4693	99.7078
SDNA129	115115	86798	82813	16569	467.2	0.4275	0.4750	99.4967
SDNA130	131747	99517	93214	16569	521.1	0.4414	0.4796	99.5903
SibA009	108290	81051	75611	16569	418.7	0.3966	0.4454	99.6505
SibA096	117913	86352	79764	16569	445.7	0.4031	0.4745	99.6398
SibYDe54	188934	84086	68056	16569	381.1	0.4150	0.4609	98.5996
SibYDy05	166306	113451	101248	16569	564.6	0.4208	0.4611	99.7657
SibYM002	128444	94400	79347	16569	439	0.4534	0.4640	99.3113
SibYO025	244055	119.6	183211	16569	325.9	0.4227	0.4376	98.7532
TXHis033	148651	107405	98198	16569	551.5	0.4256	0.4742	99.6623
TXHis117	126957	97866	91430	16569	514.6	0.4093	0.4670	99.5657
TXHis135	106114	80509	75416	16569	423.3	0.4379	0.4790	99.6202
TXHis167	91197	68407	63634	16569	360.4	0.4103	0.4660	99.8405
WANA007	142854	101578	95066	16569	519.9	0.4083	0.4716	99.6193
WANA037	124356	94550	88844	16569	499.3	0.4411	0.4765	99.6988
WANA050	138009	106268	96981	16569	547.6	0.4291	0.4734	99.6329
WANA062	123374	85425	76117	16569	427	0.4023	0.4528	99.8648
WANA065	128911	92352	85121	16569	469.9	0.4330	0.4688	99.6542
WANA093	143730	101625	95302	16569	535.9	0.4267	0.4782	99.6900

Appendix C: Phylogenetic Tree using B4a Haplogroups



Appendix D: Phylogenetic Tree using B4c Haplogroups



Appendix E: Phylogenetic Tree using B4b1 Haplogroups